UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO
DOCTORADO EN CIENCIAS BIOMÉDICAS
INSTITUTO DE ECOLOGÍA

LA DECISIÓN DEL DESTINO CELULAR COMO UNA PROPIEDAD EMERGENTE EN UN PAISAJE EPIGENÉTICO: MODELOS DINÁMICOS DE CIRCUITOS Y MÓDULOS GENÉTICOS

TESIS
QUE PARA OPTAR POR EL GRADO DE:
DOCTOR EN CIENCIAS

PRESENTA:
JOSÉ DÁVILA VELDERRAIN

TUTOR PRINCIPAL:
DRA. MARÍA ELENA ALVAREZ-BUYLLA ROCES
INSTITUTO DE ECOLOGÍA, UNAM

MIEMBROS DEL COMITÉ TUTOR:

DR. CARLOS VILLARREAL LUJÁN
INSTITUTO DE FÍSICA, UNAM

DR. HERNÁN LARRALDE RIDAURA
INSTITUTO DE CIENCIAS FÍSICAS, UNAM

MEXICO DF, AGOSTO 2015

...

Agradecimientos

Primeramente me gustaría agradecer a mi asesora *Elena* por siempre tener sus puertas abiertas para preguntas o discusiones; en especial por su apoyo, confianza, y la coordinacion de todos los proyectos de investigacion – y por siempre respaldar mis inquietudes científicas. También me gustaría agradecer a *Jorge Armando Verdusco Martínez* por ser un gran maestro, colega y amigo durante y desde la licenciatura. A *Juan Carlos Martínez García* por compartir horas de discusiones interesantes y por su colaboración en mi trabajo.

También me gustaría expresar aquí mi gratitud a aquellos autores cuyos trabajos han motivado en gran medida mi gusto por la ciencia en general, y por el enfoque de sistemas complejos a la biología en particular. Principalmente a: *Sui Huang, Stuart Kauffman y Kunihiko Kaneko* – a quienes no tengo el gusto de conocer personalmente.

Por último gracias a:
Mis co-asesores *Carlos Villarreal y Hernán Larralde.*
Mi co-asesor *Stephan Ossowski*, quien me apoyó durante una estancia en EMBL–CRG unit.
Mis colegas del Laboratorio de Genética y Evolución de Plantas.
Mi colega *Shalu Jhanwar* de EMBL–CRG unit.
Mis coautores
Mi familia

Agradezco el apoyo financiero del CONACYT.

Contents

La Decisión Del Destino Celular Como Una Propiedad Emergente En Un Paisaje Epigenético:

Modelos Dinámicos De Circuitos Y Módulos Genéticos

por

José Dávila Velderrain

Resumen

De igual manera que los humanos tomamos decisiones, las células que constituyen al humano también toman decisiones – las cuales son requeridas para producir al humano en primer lugar. Durante el desarrollo de organismos multicelulares las células deciden acerca de sus destinos mientras proliferan. A diferencia de los humanos, sin embargo, las células no tienen conciencia. ¿Cómo podemos entonces entender las decisiones sobre el destino celular durante el desarrollo como una consecuencia natural del funcionamiento celular interno? Esta tésis tiene el propósito de presentar ideas clarificadoras acerca de esta pregunta general. Particularmente, sobre como se puede explotar la claridad conceptual de un modelo metafórico de hace más de medio siglo, el Paisaje Epigenético de Waddington, con el objetivo de formular modelos mecanicistas sobre la decisión del destino celular basados en el papel organizacional de redes regulatorias genéticas subyacentes. Decisiones del destino celular resultan como una consecuencia de estos sistemas moleculares regulatorios, como tal, se espera que su papel organizacional sea un evento persistente durante la evolución. Un segundo propósito en esta tesis es estudiar la historia evolutiva y la relevancia de la conservación de una red regulatoria genética bien caracterizada y validada: la red de regulación genética del establecimiento del destino celular en la flor de *Arabidopsis thaliana*.

Métodos Se hace uso extensivo de modelos de redes regulatorias genéticas y análisis matemáticos de su dinámica. Modelos convencionales de redes regulatorias genéticas son extendidos para proponer un grupo de modelos dinámicos definido aquí como modelos del Paisaje Epigenético de Atractores. Para los análisis evolutivos se utilizan métodos estadísticos que permiten inferir el papel de distintos tipos de selección natural en linajes y sitios específicos para genes de un módulo regulatorio conservado en las plantas angiospermas.

Resultados y Conclusiones De manera global, reportamos dos resultados principales: (1) integramos propuestas de modelado necesarias para sustanciar la propuesta de que el grupo de modelos definido aquí como modelos del Paisaje Epigenético de Atractores constituyen la extensión más natural para el protocolo ya establecido de modelado de redes regulatorias genéticas, y una adición valiosa para las herramientas de la biología de sistemas. (2) Presentamos evidencia que indica que la relevancia funcional de redes regulatorias que especifican destinos celulares en la dinámica del desarrollo restringe su capacidad de sufrir un alto grado de variación durante la evolución. En otras palabras, los módulos regulatorios del desarrollo parecen ser procesos

clave que se encuentran conservados (no estń cambiando) en sistemas biológicos que presentan un procesos de desarrollo.

La Decisión Del Destino Celular Como Una Propiedad Emergente En Un Paisaje Epigenético:

Modelos Dinámicos De Circuitos Y Módulos Genéticos

by

José Dávila Velderrain

Abstract

Much as humans make decisions during their lives, the cells that constitute the human also make decisions – which are required to produce the human in the first place. During the development of a multicellular organism cells decide about their fate while proliferating. Unlike humans, however, cells do not have consciousness. How are we to understand cell-fate choices during development as a natural consequence of their inner-workings? The present thesis is meant to provide insights into this general question. Particularly, into how we can exploit the conceptual clarity of a half-century old metaphoric model, Waddington's Epigenetic Landscape, in order to derive mechanistic, post-genomic models of cell-fate decision based on the orchestrating role of underlying gene regulatory networks. Cell-fate decisions result as a natural consequence of such molecular regulatory systems, as such, their orchestrating role is expected to be a persistent event during evolution. A second major concern in this thesis is the evolutionary history and relevance of gene regulatory networks persistence.

Methods Models of gene regulatory networks and conventional mathematical analyzes of their dynamics are extensively used through the thesis. Conventional models of gene regulatory networks are extended in order to propose a group of dynamical models defined here as *Epigenetic Attractors Landscape* models. Conventional molecular evolutionary analysis are used.

Results and conclusions Overall, we report two main results: (1) we present the necessary background and modeling proposals to substantiate the claim that the group of models defined here as Epigenetic Attractors Landscape models are the most natural extension to the already established protocol of gene regulatory network modeling, and a valuable addition to the systems biology toolkit. (2) We present evidence indicating that the functional relevance of gene regulatory networks specifying cell-fates in developmental dynamics precludes them for having a high degree of variation during evolution. In other words, developmental regulatory modules seem to be key conserved, unchanging processes in biological systems undergoing development.

Keywords Gene Regulatory Networks, Epigenetic Landscape, Systems Dynamics, Epigenetic Attractors Landscape, Evolutionary Systems Biology.

Chapter 1

Introducción General

The generality of the paradox
... that the more facts we learn the less we understand the process we study ...
suggested some common fundamental flaw of how biologists approach problems.
— Yuri Lazebnik, *Can a biologist fix a radio?* (2002)

1.1 ¿Por qué estudiar la decisión del destino celular?

Este proyecto es el resultado de tres inquietudes principales que se fueron desarrollando durante mis estudios – las cuales se fueron concretizando en gran medida gracias a los antecedentes producidos y las discusiones llevadas acabo en el Laboratorio de Genética y Evolución de plantas del Instituto de Ecología. Estas inquietudes, aunque un tanto dispersas a primera vista, se relacionan dada su intersección con el problema general sobre el entendimiento del origen y la regulación del proceso de desarrollo de un organismo multicelular; particularmente, el proceso de diferenciación celular – i.e., la decisión del destino celular.

La primera inquietud se puede expresar de la siguiente manera: dado que cada célula de un organismo multicelular contiene el mismo conjunto de genes (y el mismo genoma), y considerando que todas sus células se originan de una sola, ¿cómo es que durante el desarrollo las células adquieren diferentes fenotipos celulares de manera robusta y reproducible? Y, por otro lado, ¿como es que el desarrollo de enfermedades degenerativas a edades avanzadas presenta manifestaciones fenotípicas anormales, pero estas, de forma similar, se manifiestan también de manera robusta y reproducible? Estas observaciones sugieren que existe un mecanismo subyacente que de alguna manera regula este comportamiento y que no depende directamente

de la presencia o ausencia de genes individuales.

La segunda inquietud es de naturaleza metodológica y distingue dos aspectos, uno conceptual y uno práctico. Por un lado, ¿existe algún marco teórico-conceptual que permita discutir de manera concreta los problemas expresados arriba? Por otro lado, ¿contamos con herramientas teóricas suficientes para lograr un entendimiento de las observaciones mas allá de la descripción? ¿Cómo podemos formalizar las preguntas en modelos con fines predictivos? ¿Es necesario proponer nuevas herramientas?

Por último, la tercera inquietud surge por deducción lógica a partir del supuesto enunciado en la primera inquietud: si existe un mecanismo subyacente que regula la diferenciación celular durante el desarrollo, este mecanismo debió haber surgido en etapas tempranas de la muticelularidad; por lo tanto, es razonable pensar que este mecanismo se encuentra conservado en organismos que manifiestan un proceso de desarrollo similar. ¿Existe evidencia de esto? En particular, para explorar esta ultima pregunta en la presente tesis se estudia un sistema biológico específico: el desarrollo temprano de la flor (ver abajo).

1.2 Definición del Problema de Estudio

En este proyecto las inquietudes generales expresadas en la sección anterior se definen de manera concreta y operacional de la siguiente manera.

El problema de la decisión del destino celular A nivel celular, tanto el desarrollo normal como el desarrollo de enfermedades degenerativas involucra múltiples eventos de diferenciación celular. Específicamente, una célula con la capacidad de adquirir más de un fenotipo discreto al diferenciarse, en cada evento de diferenciación adquiere solo un fenotipo. Para el propósito de este proyecto, el problema de la decisión del destino celular consiste en entender: (1) como se establecen estos fenotipos potenciales y (2) la forma en que la célula en cuestión adquiere un fenotipo y no otro. Se asume que este comportamiento resulta de manera natural a partir del funcionamiento interno de la célula que se pretende estudiar por medio del modelado matemático y computacional.

La dinámica de redes regulatorias como un mecanismo subyacente Este proyecto toma como principal hipótesis de trabajo lo siguiente: la acción concertada de los genes y sus interacciones representadas en redes regulatorias genéticas restringe el comportamiento permisible de las células, y como resultado de estas restricciones los destinos celulares potenciales son especificados. En relación a esta hipótesis, se considera que la perspectiva del comportamiento celular basada en la teoría de sistemas dinámicos brinda un marco teórico-conceptual formal y concreto que permite estudiar el problema de la toma de decisión celular de manera natural.

Modelos del *Paisaje Epigenético* asociado a redes regulatorias genéticas A pesar de contar con diversos modelos establecidos para el modelado de redes regulatorias genéticas, estos cuentan con limitaciones cuando se intenta abordar de manera natural las preguntas más relevantes para el problema de la toma de decisión celular. En linea con antecedentes recientes en el modelado de la diferenciación celular, en este proyecto se propone el uso de extensiones de modelos dinámicos de redes regulatorias genéticas con la intensión de modelar un paisaje epigenético subyacente. Los modelos resultantes permiten abordar el problema de manera natural.

Conservación evolutiva de una red regulatoria genética Motivados por la tercer inquietud sobre la existencia y conservación de un mecanismo subyacente a un proceso de desarrollo robusto, en este proyecto se plantea la hipótesis de que la red regulatoria genética caracterizada como orquestador de un proceso de diferenciación celular se encuentra conservada entre especies que manifiestan el mismo proceso. En particular, se prueba esta hipótesis utilizando como sistema de estudio el desarrollo temprano de la flor. Dado que el patrón floral en términos de tipos de órganos de la flor y organización espacio-temporal de los mismos están conservados en prácticamente todas las plantas angiospermas, originalmente se infirió la existencia de un mecanismo subyancente robusto. Se probó la existencia de tal mecanismo mediante la propuesta de una red regulatoria genética descrita originalmente en *Arabidopsis thaliana*. Considerando este antecedente, en el presente trabajo exploramos (1) si los componentes de la red regulatoria genética están conservados a nivel de secuencia en las plantas angiospermas para las cuales se ha secuenciado el genoma y (2) si existe evidencia molecular sugerente de la ocurrencia de restricciones funcionales a la evolución molecular de la red regulatoria.

1.3 Esquema General de la Tesis

La tesis está estructurada de la siguiente manera. Los *Artículos I y II* forman una parte conceptual en donde se introduce la perspectiva de la dinámica de una red regulatoria genética como un mecanismo cuyas restricciones regulatorias especifican los fenotipos celulares observables y por lo tanto los destinos celulares. Se discute esta perspectiva en el contexto del abordaje convencional de la biología molecular y la biología evolutiva. Los detalles técnicos sobre los métodos necesarios para el estudio y análisis de modelos dinámicos de redes regulatorias genéticas se presentan en los *Artículos III y IV*. El *Artículo V*, por otro lado, argumenta que el problema de la toma de decisión celular requiere de metodologías que van mas allá de los modelos convencionales de redes regulatorias genéticas. En particular, se propone la integración de modelos propuestos recientemente bajo el marco teórico de la formalización del *Paisaje Epigenético de Atractores* a partir de los modelos dinámicos de redes regulatorias genéticas como un enfoque natural para estudiar el problema. Los *Artículos III, IV y V* entonces exponen el componente metodológico principal que es aplicado en los siguientes artículos de la tesis. Cabe destacar que estos tres artículos ofrecen antecedentes importantes sobre el modelaje en general y sobre modelos dinámicos en particular, ofreciendo una introducción técnica al modelado en biología. Adicionalmente, los *Artículos III, IV y V* – basados en el lenguaje del estudio de sistemas complejos – introducen el marco teórico-conceptual necesario para abordar de manera concreta el problema de la toma de decisiones celulares.

En el resto de la tesis se presenta la aplicación a sistemas biológicos particulares del marco conceptual y los métodos desarrollados en los artículos anteriores. En el *Articulo VI* se toma uno de los modelos de redes de regulación genética mas estudiados – i.e., la red de regulación genética del establecimiento de los destinos celulares durante el desarrollo temprano de la flor de Arabidopsis – y se presenta un marco metodológico integrativo para estudiar el papel funcional de genes individuales en el contexto de la toma de decisiones celulares mediante modificaciones estructurales al Paisaje Epigenético subyacente. En el *Articulo VII* se extiende el abordaje al estudio de otro orgasnismo multicelular, el humano, y en particular a los procesos celulares durante el desarrollo de una manifestación patológica: la carcinogenesis. En este artículo se propone un nuevo modelo de red regulatoria genética como mecanismo genérico subyacente en el establecimiento de los fenotipos celulares observados durante la transformación tumorigénica de

lineas celulares epiteliales, y se analiza su Paisaje Epigenético asociado. En el *Articulo VIII* se presenta una implementación novedosa de los métodos para el modelaje del Paisaje Epigenético asociado a redes regulatorias genéticas. En el *Articulo IX*, se presenta un enfoque empírico para abordar la pregunta sobre la conservación evolutiva de un mecanismo subyacente en la determinación del destino celular. Específicamente, se toma nuevamente como sistema de estudio el proceso de desarrollo temprano de la flor, y mediante el uso de las secuencias genómicas disponibles de plantas con flor, se prueba la conservación de la red de regulación genética tanto en composición de genes como en propiedades de secuencia. Por último, en el *Articulo IX* se presenta el producto de la colaboración con la división experimental del laboratorio: se propone un nuevo modelo de red regulatoria genética y su asociado paisaje epigenético de atractores a partir de datos experimentales originales obtenidos en el laboratorio. Se muestra como tal interacción teórico-experimental permite generar una explicación mecanicista a los eventos de transición observados en el meristemo de flor.

A lo largo de la tesis nos referimos a todas las publicaciones de manera genérica como *Artćulos*, sin distinguir entre su naturaleza específica.

En resumen, los objetivos concretos del proyecto general de tesis fueron:

- Contrastar la perspectiva de un modelo conceptual de mapeo de genotipo a fenotipo uno a uno con la perspectiva de un modelo de mapeo en términos del rol auto-organizacional de redes regulatorias genéticas (Articulo I).

- Proponer el modelo del Paisaje Epigenético asociado a una GRN como un marco teórico para el estudio del efecto que tiene la generación de variación fenotípica durante el desarrollo en la evolución (Articulo II).

- Revisar y explicar los aspectos prácticos y metodologías involucradas en el planteamiento, formalización y análisis de redes regulatorias genéticas (Articulo III).

- Describir y comparar los enfoques mecanicista y descriptivo (inferencial) en el modelado de redes regulatorias genéticas, con énfasis en terminología y aspectos prácticos asociados (Articulo IV).

- Introducir el término de *Paisaje Epigenético de Atractores* como la formalización del modelo conceptual del Paisaje Epigenético de Waddington en el contexto de las redes regulatorias genéticas y la teoría de sistemas dinámicos. Revisar y discutir las estrategias de modelado del *Paisaje Epigenético de Atractores* (Articulo V).

- Proponer un marco metodológico para extender modelos de redes regulatorias genéticas con la intensión de investigar el impacto de perturbaciones a genes específicos en la toma de decisión celular como resultado de la re–estructuración del Paisaje Epigenético subyacente (Articulo VI).

- Integrar datos experimentales para proponer un modelo de red de regulación genética para el proceso de transformación tumorigénica in vitro por inmortalizacion espontanea. Mediante el análisis dinámico de la red y su Paisaje Epigenético subyacente, probar si los componentes moleculares y sus interacciones son necesarios y suficientes para recuperar los destinos celulares y transiciones observadas in-vitro e in-vivo(Articulo VII).

- Proponer una implementación novedosa de los métodos de modelaje del Paisaje Epigenético de Attractores asociado a redes regulatorias genéticas y hacerla disponible a la comunidad científica (Articulo VIII).

- Probar si los componentes de la red de regulación genética del establecimiento de los destinos celulares durante el desarrollo temprano de la flor de Arabidopsis se encuentran conservados a nivel molecular a lo largo de las plantas con flor. Probar si existe evidencia de que el módulo regulatorio ha sido sometido a restricciones funcionales durante la evolución (Articulo IX).

1.4 Información de Artículos

Articulo I: ensayo publicado en la revista *INTERdisciplina , UNAM* [Dávila-Velderrain y Álvarez-Buylla Roces].

Articulo II: capítulo publicado en el libro *Frontiers in Ecology, Evolution and Complexity, CopIt ArXives* [Davila-Velderrain *et al.*, 2014a].

Articulo III: capítulo publicado en el libro *Flower Development, Springer* [Azpeitia *et al.*, 2014].

Articulo IV: capítulo publicado en el libro Plant Functional Genomics: Methods and Protocols, Springer [Davila-Velderrain *et al.*, 2015a].

Articulo V: artículo publicado en la revista *Frontiers in Genetics - Systems Biology* [Davila-Velderrain *et al.*, 2015b].

Articulo VI: artículo en prensa en la revista *BMC Systems Biology.*

Articulo VII: artículo sometido a la revista *Journal of The Royal Society Interface.*

Articulo VIII: capítulo en preparación para ser sometido a la revista *Frontiers in Genetics - Bioinformatics and Computational Biology* [Davila-Velderrain *et al.*, 2014a].

Articulo IX: artículo publicado en la revista *Molecular Biology and Evolution* [Davila-Velderrain *et al.*, 2014b].

Articulo X: artículo publicado en la revista *Molecular Plant* [Pérez-Ruiz *et al.*, 2015].

Chapter 2

Introdución al Marco Teórico-Conceptual

...biology is finally ready for its own "theory branch"
— Arthur D Lander, *The edges of understanding* (2010)

José Dávila-Velderrain* and Elena Álvarez-Buylla Roces**

Linear Causation Schemes in Post-genomic Biology: The Subliminal and Convenient One-to-one Genotype-Phenotype Mapping Assumption

Abstract | In this essay we question the validity of basic assumptions in molecular biology and evolution on the basis of recent experimental data and through the lenses of a systems and nonlinear perspective. We focus our discussion on two well-established foundations of biology: the flow of information in molecular biology (i.e., the central dogma of molecular biology), and the "causal" linear signaling pathway paradigm. Under both paradigms the subliminal assumption of a one-to-one genotype-phenotype mapping (GPM) constitutes an underlying working hypothesis in many cases. We ask if this is empirically sustainable in post-genomic biology. We conclude that when embracing the notion of complex networks and dynamical processes governing cellular behavior—a view now empirically validated—one-to-one mapping can no longer be sustained. We hypothesize that such subliminal and sometimes explicit assumption may be upheld, to a certain degree, because it is convenient for the private appropriation and marketing of scientific discoveries. Hopefully, our discussion will help smooth the undergoing transition towards a more integrative, explanatory, quantitative and multidisciplinary systems biology. The latter will likely also yield more preventive and sustainable medical and agricultural developments, respectively, than a reductionist approach.

Keywords | post-genomic biology – genotype-phenotype mapping – genetic determinism – flow of genetic information

Introduction

SCIENCE IS MOSTLY PRACTICED out of consensus. Scientific progress, however, is also sustained by the continual challenge to accepted ideas. Unstated agreements break from time to time, and then—some say—a transition, a so-called paradigm

* Instituto de Ecología-Universidad Nacional Autónoma de México. **E-mail:** jdjosedavila@gmail.com

** Centro de Ciencias de la Complejidad-Universidad Nacional Autónoma de México. **E-mail:** eabuylla@gmail.com

Dávila-Velderrain, José and Elena Álvarez-Buylla. «Linear Causation Schemes in Post-genomic Biology: The Subliminal and Convenient One-to-one Genotype-Phenotype Mapping Assumption.» *Interdisciplina* 3, nº 5 (2015): 267-280.

shift, occurs (Kuhn 2012 [1962]). In the last decades, several authors have discussed the possibility of a paradigm shift in biology, given the apparent crisis of some of its foundational principles. (Wilkins 1996; Strohman 1997; O'Malley and Boucher 2005). In this paper, we would instead like to substantiate that a large portion of mainstream biological research subliminally embraces particular assumptions that are empirically unsustainable in this post-genomic era. Some of these assumptions are so deeply rooted that they still permeate the design, interpretation and description of a wide range of biological research at the molecular level, although, if explicitly confronted, anyone would dismiss them. Routinely we look for single, "causal" mutations responsible for complex phenotypes and assume that by finding the molecular basis of a mutation that is correlated to a particular condition, the emergence of the latter is explained. Importantly, such rationale implies that in most cases a one-to-one relationship will be possible. By extending such assumptions we define signaling pathways as autonomous entities instructing the cell how to behave under a particular condition. If pathological behavior arises, we look for the source of incorrect instructions: the mutated component or pathway. We automatically interpret any manifestation of a *learned* feature, such as drug resistance, as the consequence of the optimization principles of (*Darwinian*) adaptation by means of "random" mutation and selection. Is this recurrent bias towards *ad hoc* explanations based solely on plausibility given the evidence, or is it the mere consequence of a naively inherited tradition? We consider that an explicit presentation of some of the assumptions in light of post-genomic empirical data, and through the lenses of a systems, nonlinear perspective to biology, will clarify this question. This may prove useful for current biology students and scientists interested in multidisciplinary research.

We also include in the term post-genomic *several features that characterize modern biology: (1) abundance of experimental molecular data, (2) access to systematic ways of characterizing cellular phenotypic states, and (3) a tendency to produce quantitative data and to formulate mathematical/ computational models. Consequently, in our view,* post-genomic biology *is necessarily multidisciplinary, integrative, formal, and quantitative*

A first necessary detour: *What do we mean by post-genomic biology?* The availability of complete genome sequences (and also transcriptomes, proteomes, metabolomes, etc) obviously impacted biological research, enabling new levels of interrogation –as well as unmasking new sources of empirical support (rejection) for otherwise assumed facts. Here, however, besides access to genome-wide data, we also include in the term *post-genomic* several features that characterize modern biology: (1) abundance of experimental molecular data, (2) access to systematic ways of characterizing cellular phenotypic states, and (3) a tendency to produce quantitative data and to formulate mathematical/computational models. Consequently, in our view, *post-genomic biology* is necessarily multidisciplinary, integrative, formal, and quantitative.

The Most Basic, Naive Assumption: The One-to-One GPM

Nowadays, it is common to think about the relationship between genotypes and phenotypes in terms of some kind of complex mapping (Kauffman 1993; Mendoza and Álvarez-Buylla 1998; Wagner and Zhang 2011; Davila-Velderrain and Álvarez-Buylla 2014; Ho and Zhang 2014). The concept of a "genotype-phenotype map" can be traced back to Alberch, who elegantly proposed a model based on the principles of systems dynamics to express the inadequacy of what some call (molecular) *genetic determinism*, i.e., the assumption that genes directly determine phenotypes (Alberch 1991). Equally limited would be to assume an *epigenetic determinism*. Importantly, such A gene-centered assumption is the conceptual basis of the often invoked metaphors of a 'genetic blueprint' or a 'genetic program' (Pigliucci 2010). Furthermore, it also implies a linear relationship between genotypes and phenotypes; in other words, a *one-to-one* mapping. This simplistic model is attractive, since it naturally embraces a cause-and-effect interpretation, which makes it intuitively appealing. But if we think about this assumption of *one genotype specifically producing a particular phenotype*, we have to address how such a simplistic view can fit any observation. Nonetheless, this one-to-one model is still at the basis of most mainstream programs of biomedical or biotechnological developments (e.g., transgenic crops).

A second necessary detour: w*hat genotype and phenotype?* In the epistemology of evolution and biology, in general, it is common to talk about genotype and phenotype as absolute terms. But these can be defined at different levels, and in practice genotype and phenotype distinctions are just partial and dynamical (Lewontin 2011). In post-genomic biology this distinction is commonly aided by the use of simple GPM models (see, for example Soyer 2012). Consequently, there is not only one type of genotype and phenotype. A GPM model can be specified in different ways. For the sake of this essay we establish

that the genotype will be represented by a gene regulatory network (GRN) and the phenotypes by a gene expression profile or configuration (see below). Nevertheless, it is noteworthy that in the current era of next-generation sequencing (NGS) and single-cell biology, the empirical characterization of the complete genotypes of multiple individual cells is becoming feasible. Unfortunately, for both conceptual and technical reasons, the same cannot be said for phenotypes—although specific systematic phenotyping strategies are under development (see, for example Houle et al. 2010; Hancock 2014).

One-to-One Genotype-Phenotype Mapping and the Central Dogma

Crick declared "the central dogma of molecular biology" first in 1958 and then it was reiterated once again in 1970 (Crick 1958, 1970). In simple terms, the *dogma* posits that information flows within cells from DNA to RNA to proteins; and, as a result, the cellular phenotype is determined (Shapiro 2009). The simplifications involved in the model have been already questioned from an information viewpoint, concluding that discoveries in the last decades have made the dogma untenable (Shapiro 2009). Here we focus instead on the cemented role of the dogma regarding the implicitly assumed linear and unidirectional scheme of causation of molecular phenotypes. According to an explicit interpretation of the dogma one gene encodes for one protein, which somehow determines one observable trait (i.e. phenotype). This simplistic view can be framed effectively into a one-to-one GPM model (see Figure 1a). How do we define a phenotype? Here a phenotype is assigned to a molecule, a protein, because it is said to have a *function*. This function should be then an observable characteristic of the cell (organism). Therefore, the first one-to-one GPM to discuss would be: a gene (i.e., the genotype) codes for a protein, which performs a specific function that determines an observable characteristic (i.e., the phenotype).

Is this One-to-One (Gene-to-Function) Model Empirically Sustainable in Post-Genomic Biology?

A first difficulty that we can think of is conceptual in nature. What do we mean by *function*? Defining a function in biology is not trivial (Huang 2000; Huneman 2013; Brunet and Doolittle 2014, Doolittle *et al.* 2014). First of all, the function assignment can be given to entities at multiple levels of molecular organization; such as gene, protein, protein domain, protein complex, or pathway (Huang 2000). In the last years, researchers in the areas of genomics and epigenomics are even advocating the mapping of function at genome-wide level and single-nucleotide resolution (Kellis *et al.* 2013). For the sake of concreteness, let us

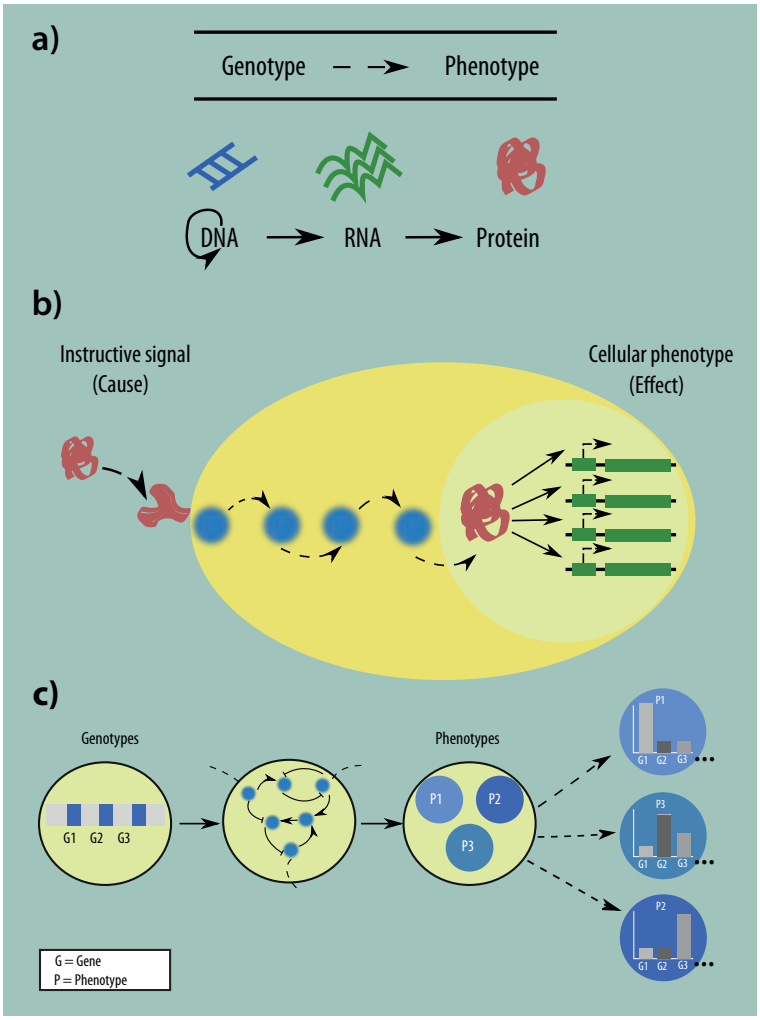


Figure 1. Schematic representation of the GPM exposed in the main text. a) One-to-one GPM model representing the central dogma of molecular biology: a gene (i.e., the genotype) codes for a protein, which performs a specific function that determines an observable characteristic (i.e., the phenotype). b) One-to-one GPM model representing the causal linear signaling pathway paradigm: genes code the proteins involved in the pathway (genotype), and these map one specific molecular signal (instruction) to a one specific cellular phenotype. c) A non-linear GPM representing cell phenotype specification by GRN dynamics: genes in a single genome (genotype) interact in complex GRNs whose regulatory interactions ultimately determine observable cell phenotypes.

just focus on function at the protein level. Although what we define as protein function is most of the times conditional on the context –i.e., cellular environment– (Huang 2000), for the purpose of our discussion, let us also assume that a protein function can be invariably assigned. Thus, in the simple one-to-one model, one gene is invariably linked to a specific function through the action of a protein.

According to the most recent assembly version of the human genome in Ensembl database (http://www.ensembl.org/), humans have 20,389 coding genes, 9,656 small noncoding genes and 14,470 long non-coding genes. A first obvious observation is that not all genes code for proteins. Two post-genomic facts: (1) most of the human genome is non-protein-coding (Alexander *et al*. 2010), and (2) transcription occurs much more often than anticipated (Carninci *et al*. 2005; Cheng *et al*. 2005). Do the genes that do not encode proteins also define a phenotype? Well, probably, in some way; but surely not by means of a one-to-one GPM, given the emerging view that non-coding transcription is tightly linked to gene regulation and cell-type specification (Natoli and Andrau 2012). For example, it was recently shown that RNA transcribed from enhancers, the so-called eRNA, is able to regulate transcription (Plosky 2014). As we will see below, gene regulation in itself is the core mechanism behind the definition of gene regulatory networks; it is also fundamental for understanding network collective behavior. Conceptualizing cell behavior in terms of molecular networks, in turn, represents a complete deviation from a one-to-one GPM (see below).

Recent assembly version of the human genome in Ensembl database, humans have 20,389 coding genes, 9,656 small noncoding genes and 14,470 long non-coding genes. A first obvious observation is that not all genes code for proteins. Two post-genomic facts: (1) most of the human genome is non-protein-coding and (2) transcription occurs much more often than anticipated. Do the genes that do not encode proteins also define a phenotype?

Besides (non)coding genes, the number of proteins coded in the human genome and represented by transcript modifications has been estimated to be between 50,000 and 500,000 (Uhlen and Ponten 2005). Considering the now known number of both genes and (estimated) proteins in other organisms, several authors have pointed out that genomic (and proteomic) complexity are not correlated with phenotypic complexity (see, for example Huang 2002). This empirical fact again is not consistent with what we would expect by extension of the dogma.

Beyond curiosity awakened by newly generated genomic data, a more serious drawback of the one-to-one GPM associated with the *central dogma* is that it completely ignores gene interactions (Tyler *et al*. 2009). Epistasis refers to the phenomenon in which the functional effect of one gene is conditional on other

genes (Phillips 2008), whereas *Pleiotropy* refers to one function being affected by multiple genes (Stearns 2010); these two phenomena are well-established facts (and concepts) in classical and modern genetics (Lehner 2011; Wagner and Zhang 2011). Nowadays such genetic interactions are being studied systematically at a genomic scale. For example, it is now possible to test millions of different combinations of double mutants and to evaluate their effects on a quantifiable function, as Costanzo and colleagues did using the budding yeast, *Saccharomyces cerevisiae* (Costanzo *et al.* 2010). Studies such as this one have clearly shown that the effect of one gene on a specific phenotype depends on the activity (or lack thereof) of many other genes. In this sense, a *genetic* interaction is defined on the base of this conditional functional effect. Although a careful discussion of epistasis and pleiotropy is beyond the scope of this paper, it is noteworthy that such mechanisms are closely related with two undeniable types of experimental evidence: (1) very different results can be produced from a nearly identical set of genes or the same genotype can produce contrasting phenotypes, and (2) virtually identical phenotypic end points can be reached by using extremely different genotypes. Evidently, these facts do not fit a one-to-one GPM. Although seemingly paradoxical, both statements can be perfectly reconciled by considering a many-to-many GPM model in which interactions among genetic and non-genetic components are explicitly considered; a view much more consistent with how living, adaptable systems behave and evolve.

One-to-One Mapping and Signaling Pathways

Extending the one-to-one view to a higher level, molecular biologists apply it to associating an altered signaling pathway to a particular phenotypic condition. Extracellular signals are transmitted by intermediary to effector proteins; which eventually activate the sets of genes responsible for the establishment of "*appropriate"* phenotypes. Note that the term pathway by itself makes reference to a group of events that occur orderly along a single *line.* Thus, in a sense, this multi-molecular model continues the *dogmatic* idea of linear, unidirectional information transfer. Thereby, in our view, it also effectively constitutes a one-to-one GPM (see figure 1b). Genes encode the proteins involved in the pathway (genotype), and these map unto one specific molecular signal (instruction) to one specific cellular phenotype. The linear property of signaling pathways also implies unidirectional cause-and-effect: a given instructional signal is thought to directly cause a phenotypic manifestation. Biologists have traditionally taken this simple pathway picture as a valid explanation at the molecular level for many cellular phenotypes. Not even a one-to-one approach to associate a network with a phenotype is valid (see below).

Is this One-To-One (Signal-to-Phenotype) Model Empirically Sustainable in Post-Genomic Biology?

Similar questions as the ones raised above can be posed here. For instance, are there enough signaling pathways for the number of possible extracelullar cues? Is there a direct, one-to-one, relationship among signals and phenotypes? If so, why do cellular phenotypes (i.e. cell types) seem to be discrete while, for example, signals carried by soluble growth factors display concentrations subject to continuous variation? And, more importantly, how and why are cellular phenotypes maintained after the signal has ceased? As we will explain below, rethinking cell behavior as the result of constraints imposed by regulatory interactions of complex molecular networks is useful to address these questions.

The genomic explosion has led to the brute-force characterization of molecular components and their interactions, which are now being integrated in large databases (Chatr-aryamontri *et al.* 2013). As expected, efforts have also tried to classify such components in genome-wide collections of signaling pathways in multiple organisms (Schaefer *et al.* 2009, Croft *et al.* 2010). What has been learned? Does the exhaustive characterization of pathways enable understanding of cellular phenotypes and their plasticity? In analogy to the failure of the pre-genomic prediction that by characterizing all the genes of an organism one will understand the genome-encoded rules instructing its behavior; listing molecular components and their interactions in pathways has only uncovered a picture that is much more complex than anticipated. But phenotypic manifestations are far from being explained by means of linear chains of molecular causation (Huang 2011)—or, in other words, of linear associations rather than explanatory models.

Decades of experimentation have shown that there is extensive crosstalk between the individually characterized signaling pathways. Accordingly, the phenomena of epistasis and pleiotropy explained above are naturally extended at the pathway level. While several different pathways can converge to specific phenotypes, one specific pathway and molecular signal can also produce different phenotypes depending on the context (Huang 2000). These observations suggest cross interactions beyond linear cascades. On the other hand, an effect similar to the one "caused" by a specific molecular signal can be produced by nonspecific stimuli or even in a stimulus-independent manner. For example, mechanical stimuli such as those induced by cell shape alterations can induce specific cell phenotypes without any molecular elicitor or genetic change (Huang 2000). On the other hand, given the intrinsic stochasticity of both extra- and intra-cellular biochemical reactions, cells in a lineage-specific manner can assume different and heritable phenotypes either in the absence of an associated genetic or environmental difference or by processing stochastic, nonspecific

environmental cues (Perkins and Swain 2009; Balázsi *et al.* 2011). These facts render a mechanistic explanation by means of the one-to-one GPM at the pathway level untenable, as well. The inevitable plasticity of cell behavior and the robustness of observed phenotypic manifestations call for an alternative explanatory model. We argue below that the formal perspective of cell behavior as an emergent property of the constraints imposed by gene regulatory networks provides an alternative view to how genotypes map unto phenotypes, providing a starting point for addressing otherwise highly complex processes.

Beyond the One-to-One GPM: A Network Dynamics Perspective

How do the two views (gene and signaling pathway to function one-to-one mapping) above stand in post-genomic, systems biology? Genes, encoded proteins, and linear signaling pathways are actually embedded in complex networks of genetic and non-genetic components which generally have various positive and negative feedback loops and dynamical behavior. We focus here on gene regulation, which is the basis for conceptualizing gene interactions, the fundamental property underlying nonlinear, gene regulatory networks. The concept of gene regulation itself, which is nothing new, is not consistent with a one-to-one GPM, because it implies that the phenotypic effect of one gene function will depend on the activity of other genes regulating it. Although explicit awareness of the fact that the genes coding for all the proteins in the cell are necessarily regulated by some other regulatory proteins, which are themselves also regulated, seems overwhelming; such realization can be succinctly represented in qualitative gene regulatory network (GRN) models. These are becoming very useful to follow and understand the concerted action of multiple interacting components.

A common working model in systems biology is that in which the genome is mapped directly to a GRN, and the cellular phenotype is represented by the activity of each of its genes, its expression pattern. Thus in a genotype-phenotype distinction based on GRN dynamics, a network represents effectively the genotype of the cell, while its associated expression profile represents its phenotype (Davila-Velderrain and Álvarez-Buylla 2014). The structure of the genome (and network) remains virtually constant through development while the cellular phenotype changes. Why are phenotypic changes observed through development in such robust and reproducible patterns?

The genomic nature of the GRN implies a physically coded structure, by means of which the network naturally constrains the permissible temporal behavior of the activity of each gene. For example, a specific gene *a* is regulated by a specific set of genes. Given the activity state of these regulators and the functional form of the regulation, each time gene *a* will be channelled to take

specific future states. This simple regulatory rule applies simultaneously to all the genes producing a self-organizing process that would inevitably lead to the establishment of only those cellular states (phenotypes), which are logically consistent with the underlying regulatory logic. Hence, the GRN imposes constraints on the behavior of the cell. The observed robustness and reproducibility of cell behavior emerges naturally as a self-organizing process. Any source of extracellular (non) specific inductive stimulus would inevitably converge to one of the phenotypic states which are logically consistent with the underlying regulatory logic of the network being considered.

The rationale briefly exposed above has been exploited to propose GRNs grounded on experimental data for understanding how cell-fate specification occurs during, for example, early flower development (See Mendoza and Álvarez-Buylla 1998; Espinosa-Soto *et al.* 2004; and an update in Sanchez-Corrales *et al.* 2010), and root stem cell patterning (Azpeitia *et al.* 2010); and it is now supported by a wealth of consolidated theoretical and experimental work (see, for example Huang *et al.* 2005; Azpeitia *et al.* 2014).

Importantly, in contrast to the assumptions implicit in the one-to-one GPM, interactions in the network are fundamental to the establishment of the phenotype, and thus the effect of a mutation on the manifested phenotype will be conditional on the network context of the gene under consideration (Davila-Velderrain *et al.* 2014). Given that the multitude of observed robust cellular phenotypic states would depend on network constraints due to gene regulatory interactions, the orchestrating role of GRNs effectively constitutes a many-to-many (*non-linear)* GPM, in which most components can, at the same time, constitute both causes and effects (Figure 1c).

Blind, Indifferent or market-oriented Biomedical and Biotechnological Research?

Notwithstanding all the evidence produced by almost two decades of post-genomic research, the subliminal presence of the over-simplified one-to-one GPM, although most of the time it is not credited, is undeniable. It is implicitly assumed as a main goal driving mainstream biomedical research that genes cause, for example, cancer; for they cause phenotypes by coding proteins (Huang 2013). This is also the case in biotechnological research, where it is acknowledged that a particular gene from one species in which a particular "function" is produced, can be readily put into another species expecting the same "function" (Vaeck *et al.* 1987). Considering that a myriad of studies search for "causal" mutations, apparently this gene-centric assumption is rarely noticed—or, alternatively, it is just ignored. Despite the huge amount of resources invested in

genome sequencing projects, such thing as a universal (causal) mutation for a degenerative disease has not been successfully identified (Huang 2013). Nevertheless, having specific molecules as candidate causal factors of particular diseases enables companies to develop new drugs for the market. Given the limited nature of the underlying simplistic one-to-one GPM, this approach is likely to fail. It may reproduce only based on its limited effectiveness—and mostly on marketing strategies—instead of deep explanations or much needed solutions. Importantly, such continuing search for potential molecular targets in therapeutics or single-gene golden bullet solutions to complex agricultural threats evidences the prevalence of the one-to-one GPM, i.e., by assuming that there is a protein for every disease or for any environmental challenge in agriculture.

The potential for therapy also complicates matters, for it may be a perfectly acceptable research goal regardless of its impact on improving understanding or on actually proving causation. Thus, it could be the case that biomedical research itself has not naturally evolved to such a naive state; it might be instead that the market driven technocentric character of modern "*science*" happens to stimulate the inheritance of old ideas that continue to be convenient—unfortunately for science, though, the rate of increase in conceptual understanding seems not to be following the fast-paced technological evolution.

To summarize, the prevailing paradigm implicitly assumes that genes determine cell behavior through a one-to-one GPM. Specifically, genes code proteins which directly determine phenotypes, and consequently, mutations in the genes should by themselves alter phenotypes. Therefore, targeting altered proteins produced from mutated genes seems to be the best strategy to "correct" a pathological phenotype—the same can be said of epigenetic alterations, altered pathways or even networks. However, a multitude of post-genomic evidence makes the one-to-one GPM untenable. In contrast, a GPM in terms of the orchestrating role of molecular regulatory networks, which constitutes a many-to-many GPM, naturally explains paradoxical observations and provides a formal framework for the interpretation of ever-growing post-genomic molecular data. ld

Acknowledgements

This work was supported with ERAB grants: Conacyt (Mexico) 180098 and 180380; and UNAM-DGAPA-PAPIIT: IN203113.

References

Alberch, P. «From genes to phenotype: dynamical systems and evolvability.» *Genetica* 84, nº 1 (1991): 5-11.

Alexander, R. P., G. Fang, J. Rozowsky, M. Snyder and M. B. Gerstein. «Annotating non-coding regions of the genome.» *Nature Reviews Genetics* 11, nº 8 (2010): 559-571.

Azpeitia, E., J. Davila-Velderrain, C. Villarreal and E. Álvarez-Buylla. «Gene regulatory network models for floral organ determination.» *Flower Development* (2014): 441-469.

———, M. Benítez, I. Vega, C. Villarreal and E. Álvarez-Buylla. «Single-cell and coupled GRN models of cell patterning in the Arabidopsis thaliana root stem cell niche.» *BMC systems biology* 4, nº 1 (2010): 134.

Balázsi, G., Van Oudenaarden, A. and J. J. Collins. «Cellular decision making and biological noise: from microbes to mammals.» *Cell* 144, nº 6 (2011): 910-925.

Brunet, T. D. and W. F. Doolittle. «Getting "function" right.» *Proceedings of the National Academy of Sciences* 111, nº 33 (2014): E3365-E3365.

Carninci, P., *et al*. «The transcriptional landscape of the mammalian genome.» *Science* 309, nº 5740 (2005): 1559-1563.

Chatr-aryamontri, A., *et al*. «The BioGRID interaction database: 2013 update.» *Nucleic acids research* 41 nº D1 (2013): D816-D823.

Cheng, J., *et al*. «Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution.» *Science* 308, nº 5725 (2005): 1149-1154.

Costanzo, M., *et al*. «The genetic landscape of a cell.» *Science* 327, nº 5964 (2010): 425-431.

Crick, F. H. «Central dogma of molecular biology.» *Nature* 227, nº 5258 (1970): 561-563.

———. «On protein synthesis.» *Symposia of the Society for Experimental Biology* 12 (1958): 138.

Croft, D., *et al*. «Reactome: a database of reactions, pathways and biological processes.» *Nucleic acids research*, (2010): gkq1018.

Davila-Velderrain, J., A. Servin-Marquez and E. Álvarez-Buylla. «Molecular evolution constraints in the floral organ specification gene regulatory network module across 18 angiosperm genomes.» *Molecular biology and evolution* 31, nº 3 (2014): 560-573.

——— and E. Álvarez-Buylla. «Bridging genotype and phenotype.» In *Frontiers in Ecology, Evolution and Complexity*, edited by Octavio Miramontes, Alfonso Valiente-Banuet and Mariana Benítez. CopIt ArXives, 2014.

Doolittle, W. F., T. D. Brunet, S. Linquist and T. R. Gregory. «Distinguishing between "function" and "effect" in genome biology.» *Genome biology and evolution* 6, nº 5 (2014): 1234-1237.

Espinosa-Soto, C., P. Padilla-Longoria and E. Álvarez-Buylla. «A gene regulatory network model for cell-fate determination during Arabidopsis thaliana

flower development that is robust and recovers experimental gene expression profiles.» *The Plant Cell Online* 16, nº 1 (2004): 2923-2939.

Hancock, J. M. (Ed.). *Phenomics.* CRC Press, 2014.

Ho, W. C. and J. Zhang. «The Genotype-Phenotype Map of Yeast Complex Traits: Basic Parameters and the Role of Natural Selection.» *Molecular biology and evolution* 31, nº 6 (2014): 1568-1580.

Houle, D., D. R. Govindaraju and S. Omholt. «Phenomics: the next challenge.» *Nature Reviews Genetics* 11, nº 12 (2010): 855-866.

Huang, S., G. Eichler, Y. Bar-Yam and D. E. Ingber. «Cell fate sas high-dimensional attractor states of a complex gene regulatory network.» *Physical Review Letters* 94, nº 12 (2005): 128701.

———. «Genetic and non-genetic instability in tumor progression: link between the fitness landscape and the epigenetic landscape of cancer cells.» *Cancer and Metastasis Reviews* 32, nº 3-4 (2013): 423-448.

———. «Rational drug discovery: what can we learn from regulatory networks?» *Drug discovery today* 7, nº 20 (2002): s163-s169.

———. «Systems biology of stem cells: three useful perspectives to help overcome the paradigm of linear pathways. Philosophical Transactions of the Royal Society B.» *Biological Sciences* 366, nº 1575 (2011): 2247-2259.

———. «The practical problems of post-genomic biology.» *Nature biotechnology* 18, nº 5 (2000): 471-472.

Huneman, P. *Functions: selection and mechanisms.* Springer, 2013.

Kellis, M., *et al.* «Defining functional DNA elements in the human genome.» *Proceedings of the National Academy of Sciences* 111, nº 17 (2014): 6131-6138.

Kuhn, T. S. *The structure of scientific revolutions.* University of Chicago Press, 2012 [1962].

Lehner, B. «Molecular mechanisms of epistasis within and between genes.» *Trends in Genetics* 27, nº 8 (2011): 323-331.

Lewontin, R. «The genotype/phenotype distinction.» In *Stanford Encyclopedia of Philosophy*. 2011.

Mendoza, L. and E. Álvarez-Buylla. «Dynamics of the genetic regulatory network for arabidopsis thaliana flower morphogenesis.» *Journal of Theoretical Biology* 193, nº 2 (1998): 307-319.

Natoli, G. and J. C. Andrau. «Noncoding transcription at enhancers: general principles and functional models.» *Annual review of genetics* 46 (2012): 1-19.

O'Malley, M. A. and Y. Boucher. «Paradigm change in evolutionary microbiology.» *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, 2005: 183-208.

Perkins, T. J. and P. S. Swain. «Strategies for cellular decision-making.» *Molecular systems biology* 5, nº 1 (2009).

Phillips, P. C. «Epistasis-the essential role of gene interactions in the structure and evolution of genetic systems.» *Nature Reviews Genetics* 9, nº 11 (2008): 855-867.

Pigliucci, M. «Genotype-phenotype mapping and the end of the 'genes as blueprint'metaphor.» *Philosophical Transactions of the Royal Society B: Biological Sciences* 365, nº 1540 (2010): 557-566.

Plosky, Brian S. *eRNAs Lure NELF from Paused Polymerases.Molecular Cell.* 2014.

Rose, M. R. and T. H. Oakley. «The new biology: beyond the Modern Synthesis.» *Biology direct* 2, nº 1 (2007): 30.

Sanchez-Corrales, Y. E., E. Álvarez-Buylla and L. Mendoza. «The Arabidopsis thaliana flower organ specification gene regulatory network determines a robust dif-ferentiation process.» *Journal of Theoretical Biology* 264, nº 3 (2010): 971-983.

Schaefer, C. F., *et al*. «PID: the pathway interaction database.» *Nucleic acids research*, 2009: D674-D679.

Shapiro, J. A. «Revisiting the central dogma in the 21st century.» *Annals of the New York Academy of Sciences* 1178, nº 1 (2009): 6-28.

Soyer, O. S. (Ed.). *Evolutionary systems biology* 751 Spring 2012.

Stearns, F. W. «One hundred years of pleiotropy: a retrospective.» *Genetics* 186, nº 3 (2010): 767-773.

Strohman, R. C. «The coming Kuhnian revolution in biology.» *Nature biotechnology* 15, nº 3 (1997): 194-200.

Stuart A., Kauffman. *The origins of order: Self-organization and selection in evolution.* Oxford, UK: Oxford University Press, 1993.

Tyler, A. L., F. W. Asselbergs, S. M. Williams and J. H. Moore. «Shadows of complexity: what biological networks reveal about epistasis and pleiotropy.» *Bioessays* 31, nº 2 (2009): 220-227.

Uhlen, M. and F. Ponten. «Antibody-based proteomics for human tissue profiling.» *Molecular & Cellular Proteomics* 4, nº 4 (2005): 384-393.

Vaeck, M., *et al*.«Transgenic plants protected from insect attack.» *Nature* 328 (1987): 33-37.

Wagner, G. P. and J. Zhang. «The pleiotropic structure of the genotype-phenotype map: the evolvability of complex organisms.» *Nature Reviews Genetics* 12, nº 3 (2011): 204-213.

Wilkins, A. S. «Are there 'Kuhnian'revolutions in biology?» *BioEssays* (1996): 695-696.

Bridging the Genotype and the Phenotype: Towards An Epigenetic Landscape Approach to Evolutionary Systems Biology

Davila-Velderrain J[1,2,*], Alvarez-Buylla ER[1,2,*]

[1] Instituto de Ecología, Universidad Nacional Autónoma de México, Cd. Universitaria, México, D.F. 04510, México
[2] Centro de Ciencias de la Complejidad (C3), Universidad Nacional Autónoma de México, Cd. Universitaria, México, D.F. 04510, México
[*] E-mail: jdjosedavila@gmail.com, eabuylla@gmail.com

Abstract

Understanding the mapping of genotypes into phenotypes is a central challenge of current biological research. Such mapping, conceptually represents a developmental mechanism through which phenotypic variation can be generated. Given the nongenetic character of developmental dynamics, phenotypic variation to a great extent has been neglected in the study of evolution. What is the relevance of considering this generative process in the study of evolution? How can we study its evolutionary consequences? Despite an historical systematic bias towards linear causation schemes in biology; in the post-genomic era, a systems-view to biology based on nonlinear (network) thinking is increasingly being adopted. Within this view, evolutionary dynamics can be studied using simple dynamical models of gene regulatory networks (GRNs). Through the study of GRN dynamics, genotypes and phenotypes can be unambiguously defined. The orchestrating role of GRNs constitutes an operational *non-linear* genotype-phenotype map. Further extension of these GRN models in order to explore and characterize an associated Epigenetic Landscape enables the study of the evolutionary consequences of both genetic and non-genetic sources of phenotypic variation within the same coherent theoretical framework. The merging of conceptually clear theories, computational/mathematical tools, and molecular/genomic data into coherent frameworks could be the basis for a transformation of biological research from mainly a descriptive exercise into a truly mechanistic, explanatory endeavor.

Introduction

The mechanistic understanding of the mapping of genotypes into phenotypes is at the core of modern biological research. During the lifetime of an individual, a developmental process unfolds, and the observed phenotypic characteristics are consequently established. As an example, a given individual may or may not develop a disease. Can we explain the observed outcome exclusively in terms of genetic differences and an unidirectional,

linear relationship between genotype and phenotype? Researchers in biology have mostly assumed so. Over the last decades, scientists under the guidance of such genetic-causal assumption have struggled with inconsistent, empirical observations. The biological relevance of the phenotypic variability produced during the developmental process itself, and not as the consequence of genetic mutations, has only recently started to be acknowledged [1–5].

Understanding the unfolding of the individuals phenotype is the ultimate goal of developmental biology. Evolutionary biology, on the other hand, is largely concerned with the heritable phenotypic variation within populations and its change during long time periods, as well as the eventual emergence of new species. Historically, population-level models seek to characterize the distribution of genotypic variants over a population, considering that genetic change is a direct indicator of phenotypic variation. Certain assumptions are implicit to such reasoning. Are those assumptions justifiable in light of the now available molecular data and the recently uncovered molecular regulatory mechanisms? What is the relevance of considering the generative developmental sources of phenotypic variation in the study of evolution? The aim of this paper is to highlight how a systems view to biology is starting to give insights into these fundamental questions. The overall conclusion is clear: an unilateral *gene-centric* approach is not enough. Evolution and development should be integrated through experimentally supported mechanistic dynamical models [6–13].

In the sections that follow, we first present a brief historical overview of evolutionary biology and the roots of a systematic bias towards linear causation schemes in biology. Then, we discuss the assumptions implicit in the so-called neo-Darwinian Synthesis of Evolutionary Biology – the conventional view of evolution. In the last section, we briefly describe an emerging research program which aims to go beyond the conventional theory of evolution, focusing on a nonlinear mapping from genotype to phenotype through the restrictions imposed by the interactions in gene regulatory networks (GRNs) and its associated epigenetic landscape (EL). Overall, this contribution attempts to outline how the orchestrating role of GRNs during developmental dynamics imposes restrictions and enables generative properties that shape phenotypic variation.

Darwin's Legacy

Darwin eliminated the need for supernatural explanations for the origin and adaptations of organisms when he put evolution firmly on natural grounds [14]. In the mid-19th century, Darwin published his theory of natural selection [15]. He proposed a natural process, the gradual accumulation of variations sorted out by natural selection, as an explanation for the shaping and diversity of organisms. This insight was what put the study of evolution within the realms of science in the first place [14]. Although it has had its ups and downs [16], the Darwinian research tradition predominates in modern evolutionary biology. Much of its success is due to a new (gene-centric) interpretation, the so-called neo-Darwinian modern synthesis [17]: the merging of mendelian genetics and Darwin's theory of natural selection due to prominent early 20th century statisticians. In this framework, development was left outside, and evolution is seen as a change in the genotypic constitution of a population over time. Genes map directly into phenotypes (see

Figure 1a), implicitly assuming that genetic mutation is the prime cause of phenotypic variation. Observed traits are generally assumed to be the result of adaptation, the process whereby differential fitness (the product of the probability of reproduction and survival) due to genetic variation in a particular environment, leads to individuals better able to live in such an environment.

From Natural Selection to Natural Variation

Natural selection - a force emanating from outside the organism itself - is the conceptual core of the Darwinian research tradition. Conceptually, the general process is as follows. *Random* mutations occur during reproduction; these mutations are responsible for generating different (genetic) types of individuals. The selection process then results from the fact that each type has certain survival probability and/or is able to achieve certain reproductive performance given the environment. Through this differential rate, some types are maintained while others are dismissed. It is said that, in this way, selection makes a "choice" [18]. From a wider perspective, it is generally accepted that selection is a generic process not restricted to biological evolution [19]. Any error-prone communication process in which information is consequently transmitted at different rates leads itself to a selection mechanism. However, despite the appealing conceptual clarity of the selection mechanism, it is not generally appreciated that the complexity inherent to biological systems hinders the mechanistic understanding of biological evolution. Because the reproductive performance of a given type of variant is, mainly, a function of its phenotype; the paradigmatic selection process described above is plausible when one assumes a straightforward causation of phenotype by genotype [10]. A more faithful model of biological evolution should explicitly consider a genotype-phenotype (GP) map [20, 21], a developmental mechanism which specifies how phenotypic variation is generated (Figure 1b). The generated variation is then what triggers selection [22]. Importantly, a deviation from a linear causation view of development would potentially impact the rate and direction of evolution [8, 23, 24].

Although not always discussed, Darwin himself devoted much more attention to variation than to natural selection, presumably because he knew that a satisfactory theory of evolutionary change requires the elucidation of the causes and properties of variation [25]. After all, natural selection would be meaningless without variation. Ironically, given the success of the neo-Darwinian framework, phenotypic variation to a great extent has been neglected in the study of evolution [26]. The mechanistic understanding of the sources of phenotypic variation constitutes a fundamental gap in conventional evolutionary theory. Neither Darwin, nor the founders of the neo-Darwinian modern synthesis were able to address this problem given the biological knowledge available at the time. Moreover, deviations from the basic assumptions of the conventional theory were not always generally appreciated [27].

Implicit Assumptions in Evolution

Being the development of science an evolutionary process itself, it is reasonable to expect that social-historical contingency has profoundly biased the pathways of scientific inquiry. This seems to be the case in the history of biology. For example, (1) Darwin's war against

divine explanations for biological complexity caused within the scientific community an automatic rejection for any goal-oriented activity within organisms. This situation favored the adoption of the the idea of random (uniform) variation [28, 29]. (2) The mainstream focus of neo-Dawinism on optimizing reproductive success (fitness) by natural selection of random variants; on the other hand, implicitly neglected the relevance of gene interactions (see Figure 1a) [30]. Finally, (3) the establishment of the central dogma of molecular biology (gene $\rightarrow$ mRNA $\rightarrow$ protein) further cemented a linear, unidirectional scheme of causation of molecular traits (one gene - one protein, one trait) [10]. These events are thought to be associated with a deeply rooted systematic bias towards linear causation schemes in biology [10, 31]. They also favored the adoption of three major implicit assumptions upon which the neo-Darwinian tradition was developed, namely: (1) mutational events occur randomly (e.g. unstructured) along the genome; (2) given that the phenotypic effects of successive mutations in evolution are of additive nature, gene interactions and their phenotypic influence can be, to a large extent, ignored; and (3) the phenotypic distribution of mutational effects mirrors the genetic distribution of mutations [30].

Scientists are now re-examining the most basic assumptions about evolution in light of post-genomic, systems biology [28, 32]. Compelling evidence has been presented even against assumption (1) above. For example, Shapiro has shown how a truly random (unstructured) nature of mutational events is empirically unsustainable. He has coined the term "natural genetic engineering", referring to the known operators that produce genomic changes and which are subjected to cellular regulatory regimes of epigenetic character [29]. It seems that the generative properties of genetic variation are nonuniform, and thus, biased as well. Assumptions (2) and (3) above are, instead, mainly concerned with how phenotypic variation is generated given a genetic background; or in other words, with the mechanistic understanding of the GP map. Here, we are concerned with this developmental process and its evolutionary relevance.

From Genes to Networks

At the beginning of the 21th century, biology confronted an uncomfortable fact: despite the increasing availability of whole-genome sequence data, it was not possible to predict, or even clarify, phenotypic observations. In fact, we now know that there is not sufficient information in the linear DNA sequences of the complete genomes to recover and/or understand the diverse phenotypic states of an organism. It was clear that cell behavior was much more complex than anticipated. Since then, biological research has increasingly been oriented towards a systems-level approach that goes beyond obtaining and describing large data sets at the genomic, transcriptomic, proteomic or metabolomic levels. An assumption of such *systems* approach to biology is that cell behavior can be understood in terms of the dynamical properties of the involved molecular regulatory networks. Modern molecular evolutionary studies are starting to incorporate this network thinking: genes are not individual entities upon which evolutionary forces act independently. Evolutionary forces, functional constraints, and molecular interactions are conditionally dependent on the systems level [33]. How a systems-view impacts our understanding of the GP map?

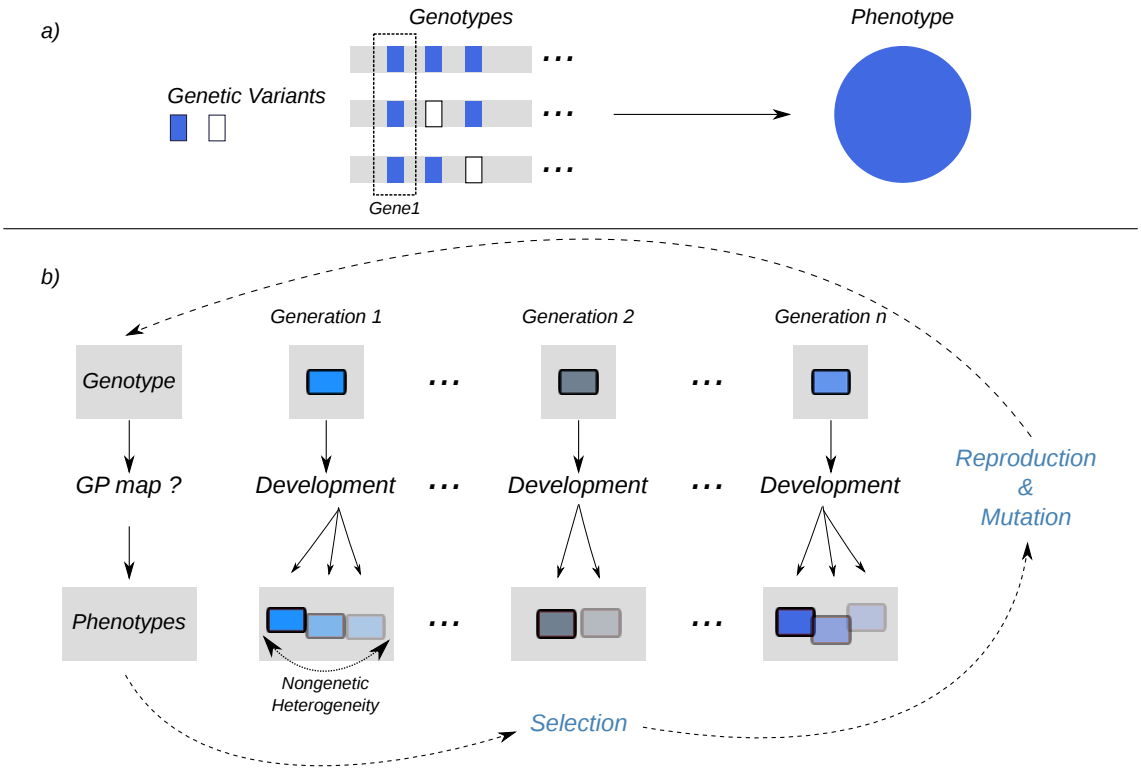


Figure 1. a) A straightforward genotype-phenotype relationship: the genetic distribution of the observed locus would completely mirror the phenotypic distribution; gene interactions are ignored; as a result, three different genotypes would correspond to the same phenotype given the locus under observation. b) A developmental process from genotype to phenotype, a GP map: through the development of an individual nongenetic phenotypic variation is generated each generation; in an evolutionary time-scale, evolution operations (blue) produce genetic variation. Selection acts on phenotypes; phenotypic variation is the product of both genetic mutational operations and epigenetic developmental processes.

Fundamental Sources of Natural Variation

Although the concepts of genotype and phenotype are fundamental to evolution, it is not straightforward to operationally define them: In practice genotype and phenotype distinctions are just partial [34]. This is partially the reason why simple theoretical models are so important for the epistemology of evolution. A common working model in systems biology is that in which the phenotypic state is defined at the cellular level. The cellular phenotype is represented by the activity of each of its genes, its expression pattern. Since the regulatory interactions among the genes within the cell constitute a network, the network effectively represents the genotype of the cell, while its associated expression profile represents its phenotype (Figure 2). The structure of the former derives directly from the genome, while the latter changes through development. In practice, we just observe certain expression patterns (e.g cell-types) - with small deviations - and not others. Why is that?

GRN developmental dynamics generates phenotypic nongenetic (epigenetic) heterogeneity

When thinking in terms of a genotype-phenotype distinction based on GRN dynamics, it is natural to consider an abstract space where all the virtually possible phenotypes reside. We call this space the *state-space*. Empirical observations suggest that something should be maintaining cells within specific, restricted regions of this space. The structured nature of the underlying GRN determines a trajectory in this state-space: given the state of the genes regulating a gene i, and the functional form of the regulation, the gene i is canalized to take specific future states. Eventually, this self-organizing process would inevitably lead to the establishment of those states which are logically consistent with the underlying regulatory logic. In this way, the GRN imposes constraints to the behavior of the cell. The resultant states are denominated *attractors* and correspond to observable cell-types. These are the basis of the well developed dynamical-systems theory of cell biology (for a review, see [35, 36]). This theory was first applied to propose a GRN grounded on experimental data for understanding how cell-fate specification occurs during early flower development (see, [37, 38] and update in [39]). Originally, the approach was inspired by theoretical work in randomly assembled networks by Stuart Kaufman [40]. In the last decades, the theory has been supported by a wealth of consolidated theoretical and experimental work (see, for example [7, 13, 41]).

Through GRN dynamics, development generates cellular phenotypes. The general acceptance of this generative role necessarily implies deviations from the neo-Dawinian framework. Importantly, (1) the effect of a perturbation (mutational or otherwise) on the manifested phenotype is not uniformly distributed (truly random) across all the genes in the network, and (2) the interactions in the network are fundamental to the establishment of the phenotype. The orchestrating role of GRNs constitutes a *non-linear* GP map: phenotypic variation does not scale proportionally to genotypic variation; it is not linear (Figure 2). Two important consequences of these mechanistic view of developmental dynamics have been eloquently pointed out recently. First, the nonlinear character of this mapping ensures that the exact same genotype (network) is able to produce several phenotypes (attractors) [40]. Second, given that molecular regulatory events are stochastic in nature, a cell is able to explore the state-space by both attracting and dispersing forces -

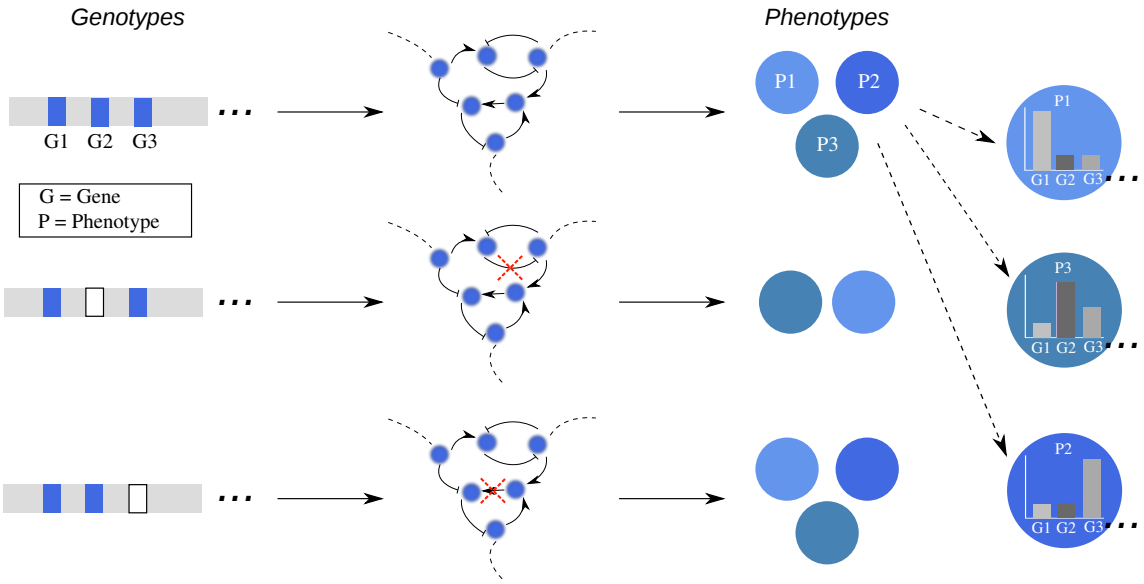


Figure 2. The orchestrating role of GRNs constitutes a *non-linear* GP map. Through the restrictions imposed by the interactions in GRNs, cellular phenotypes (represented by expression profiles) are generated. Due to the nonlinear character of GRN dynamics, the GP map is one-to-many. The effect of mutations in the phenotype is not uniformly distributed over the genes, but depends on the interactions: mutations can or cannot result in different phenotypes depending on the genetic background and the location of the affected genes in the network.

forces that slightly deviate the dynamics from the determined trajectory. Any phenotype of a cellular population at any given time is statistically distributed [10]. These sources of variation are the natural product of developmental dynamics. Consequently, at any given time, a population can manifest phenotypic variation that is relevant to evolution (heritable) in the absence of genetic variation. How can we study evolution without ignoring the fundamental role of developmental dynamics?

Evolutionary Systems Biology Approaches

A systems view to evolutionary biology, in which network models as GP mappings are considered explicitly, is under development (see, for example [9, 11, 42]). Within this general framework, several specific approaches are proposed in order to study the evolutionary consequences of considering developmental sources of phenotypic variation. In this section, we briefly present a preview of an emerging complementary approach.

Epigenetic(Attractors) Landscape Evolution

In 1950s, C.H. Waddington proposed the conceptual model of the epigenetic landscape (EL), a visionary attempt to synthesize a framework that would enable an intuitive discussion about the relationship between genetics, development, and evolution [43]. His reasoning was based on the consideration of a fact: the physical realization of the informa-

tion coded in the genes - and their interactions - imposes developmental constraints while forming an organism. Now, in the post-genomic era, a formal basis for this metaphorical EL is being developed in the context of GRNs [10, 44, 45]. The key for this formalization is an emergent ordered structure embedded in the state-space, the attractors landscape (AL). As well as generating the cellular phenotypic sates (attractors), the GRN dynamics also partitions the whole state-space in specific regions and restricts the trajectories from one state to another one. Each region groups the cellular states that would eventually end up in a single, specific attractor. These sub-spaces are denominated the attractor's *basin* of attraction. Given this (second) generative property of GRN dynamics, the formalization of the EL in this context is conceptually straightforward: the number, depth, width, and relative position of these basins would correspond to the hills and valleys of the metaphorical EL. We refer to this structured order of the basins in state-space as the AL (see Figure 3). The characterization of an AL would correspond, in practical terms, to the characterization of an EL. Is this formalized EL useful for the mechanistic understanding of phenotype generation?

Multicellular morphogenetic processes unfold naturally in the EL

The structured EL is a generative property of the GRN dynamics, but at the same time, it also constrains the behavior of a developing system. While a developing system is following its dynamically constrained trajectory in state-space, developmental perturbations from internal or external origin can deviate it. In a cellular population, then, the probability of one phenotypic transition or another during development, as well as the stationary distribution of phenotypes, would be conditioned on both the localization of the individual cells in the EL and on the landscape's structure. As a general result of this interplay, determinism and stochasticity are reconciled, and robust morphogenetic patterns can be established by a hierarchy of cellular phenotypic transitions (see, for example [44, 45]). In this way, morphogenetic processes effectively unfold on ELs. How could this theoretical framework improve the understanding of evolutionary dynamics?

We have an effective nonlinear GP map from GRN to EL. Given an experimentally characterized GRN, the EL associated to real, specific developmental processes can be analyzed ([13, 44, 45]). Both cellular phenotypes (attractors) and morphogenetic patterns are linked to the structure of the EL. Can we describe this structure quantitatively? How robust is the structure to genetic (network) mutation? Can we describe quantitatively the change in structure in response to both mutational and developmental perturbations? How slower is this rate of change in comparison to the time-scale of developmental dynamics (landscape explorations)? What are the phenotypic consequences of different relative rates of change? Does the resultant evolutionary trajectory of the reshaped EL structure subjected to mutations predicts the probability of phenotypic change (innovation) - based, for example, in the appearance of new cellular phenotypes or morphogenetic patterns? (Figure 3). Insight into these and similar questions could enhance the mechanistic understanding of the evolution of morphogenetic processes.

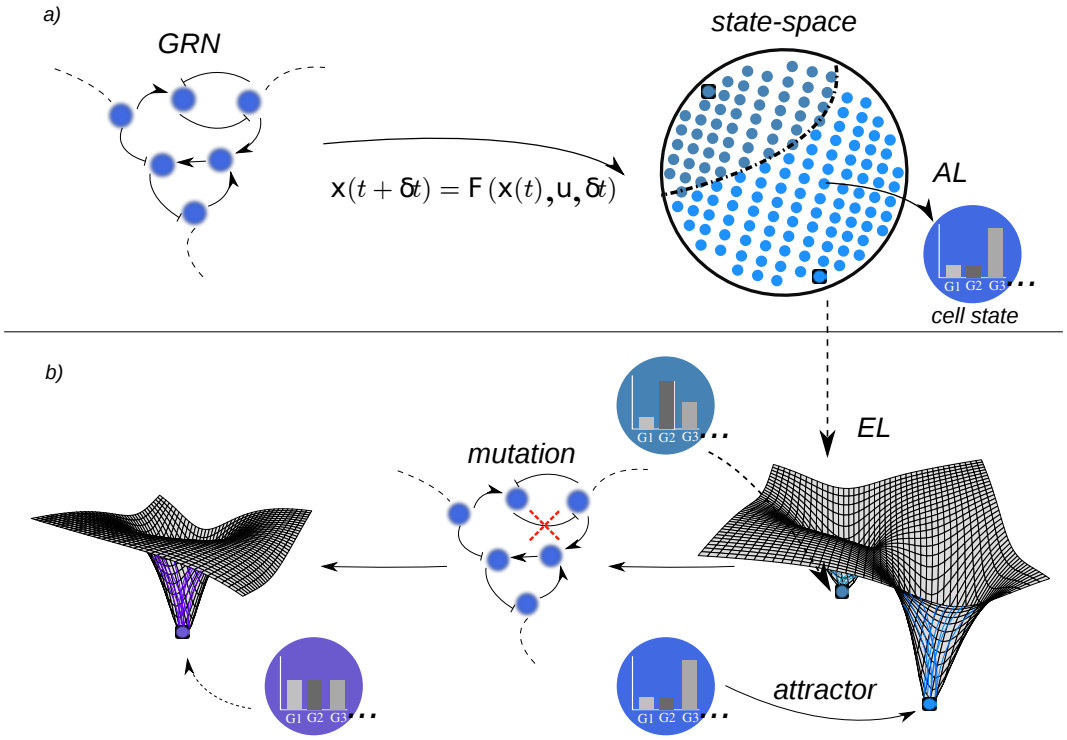


Figure 3. The Epigenetic (Attractors) Landscape. a) Through a dynamical mapping - a mathematical representation of the gene regulatory logic - GRNs generate both the cellular phenotypes (attractors) and the ordered structure of the state space - the AL. Through the structure of the AL, the EL is formalized in the context of GRNs. b) The number, depth, width, and relative position of attractors correspond to the hills and valleys of the EL. The topography of the landscape can change in response to perturbations. Mutations could eventually reshape the EL and consequently eliminate and/or generate novel phenotypes.

Conclusion and Challenges

A modern systems view to biology enables tackling foundational questions in evolutionary biology from new angles and with unprecedented molecular empirical support. Little is known about the mechanistic sources of phenotypic variation and its impact to evolutionary dynamics. The explicit consideration of these processes in evolutionary models directly impacts our thinking about evolution. Simple, generic dynamical models of GRNs, where genotypes and phenotypes can be unambiguously defined, are well-suited to rigorously explore the problem. Further extension of these models in order to explore and characterize the associated EL enables the study of the evolutionary consequences of both genetic and non-genetic sources of phenotypic variation within the same coherent theoretical framework. The network-EL approach to evolutionary dynamics is promising, as it directly manifests the multipotency associated with a given genotype. Although conceptually clear and well-founded, its practical implementation implies several difficulties, nonetheless; specially in the case of high dimensional systems. Work has been done in which the landscape associated with a specific, experimentally characterized GRN is described quantitatively in terms of robustness and state transition rates [46], for example. However, neither the methodology to derive ELs from GRNs, nor the quantitative description of ELs are standard procedures. Most approaches require approximations and are technically challenging for the case of networks with more than 2 nodes. Further research in the quantitative description of experimentally grounded GRNs is still needed in order to explore the constraints and the plasticity of ELs associated with a genotypic (network) space. In this regard, discrete dynamical models are promising tools for the exhaustive characterization of the EL, and for the study of multicellular development [45]. A second major challenge is the generalization of GRN dynamical models in order to include additional sources of constraint during development. Tissue-level patterning mechanisms such as cell-cell interactions; chemical signaling; cellular growth, proliferation, and senescence; inevitably impose physical limitations in terms of mechanical forces which in turn affect cellular behavior. Although some progress has been presented in this direction [47, 48], the problem certainly remains open.

The post-genomic era of biology is starting to show that old metaphors such as Waddington's EL are not just frameworks for the conceptual discussion of complex problems. The merging of conceptually clear theories, computational/mathematical tools, and molecular/genomic data into coherent frameworks could be the basis for a much needed transformation of biological research from mainly a descriptive exercise into a truly mechanistic, explanatory and predictive endeavor - EL models associated with GRNs being a salient example.

References

1. Feinberg AP, Irizarry RA (2010) Stochastic epigenetic variation as a driving force of development, evolutionary adaptation, and disease. Proceedings of the National Academy of Sciences 107: 1757–1764.

2. Frank SA, Rosner MR (2012) Nonheritable cellular variability accelerates the evolutionary processes of cancer. PLoS biology 10: e1001296.

3. Freund J, Brandmaier AM, Lewejohann L, Kirste I, Kritzler M, et al. (2013) Emergence of individuality in genetically identical mice. Science 340: 756–759.

4. Huang S (2009) Non-genetic heterogeneity of cells in development: more than just noise. Development 136: 3853–3862.

5. Pisco AO, Brock A, Zhou J, Moor A, Mojtahedi M, et al. (2013) Non-darwinian dynamics in therapy-induced cancer drug resistance. Nature communications 4.

6. Alvarez-Buylla ER, Azpeitia E, Barrio R, Benítez M, Padilla-Longoria P (2010) From abc genes to regulatory networks, epigenetic landscapes and flower morphogenesis: making biological sense of theoretical approaches. Seminars in cell & developmental biology 21: 108–117.

7. Jaeger J, Crombach A (2012) Lifes attractors. In: Evolutionary Systems Biology, Springer. pp. 93–119.

8. Jaeger J, Irons D, Monk N (2012) The inheritance of process: a dynamical systems approach. Journal of Experimental Zoology Part B: Molecular and Developmental Evolution 318: 591–612.

9. Wagner A (2011) The origins of evolutionary innovations. Oxford University Press, Oxford.

10. Huang S (2012) The molecular and mathematical basis of waddington's epigenetic landscape: A framework for post-darwinian biology? Bioessays 34: 149–157.

11. Soyer OS (2012) Evolutionary systems biology, volume 751. Springer.

12. Benítez M, Azpeitia E, Alvarez-Buylla ER (2013) Dynamic models of epidermal patterning as an approach to plant eco-evo-devo. Current opinion in plant biology 16: 11–18.

13. Azpeitia E, Davila-Velderrain J, Villarreal C, Alvarez-Buylla ER (2014) Gene regulatory network models for floral organ determination. In: Flower Development, Springer. pp. 441–469.

14. Ayala FJ (2007) Darwin's greatest discovery: design without designer. Proceedings of the National Academy of Sciences 104: 8567–8573.

15. Darwin C (1859) On the origins of species by means of natural selection. London: Murray.

16. Depew DJ, Weber BH (1995) Darwinism evolving: Systems dynamics and the genealogy of natural selection. Bradford Books/MIT Press.

17. Huxley J, et al. (1942) Evolution. the modern synthesis. Evolution The Modern Synthesis .

18. Nowak MA (2006) Evolutionary dynamics: exploring the equations of life. Harvard University Press.

19. Schuster P (2008) Boltzmann and evolution: some basic questions of biology seen with atomistic glasses. Boltzmanns Legacy : 217–241.

20. Lewontin RC (1974) The genetic basis of evolutionary change, volume 560. Columbia University Press New York.

21. Alberch P (1991) From genes to phenotype: dynamical systems and evolvability. Genetica 84: 5–11.

22. Schaper S, Louis A (2014) The arrival of the frequent: How bias in genotype-phenotype maps can steer populations to local optima. PloS one 9: e86635.

23. Gould SJ, Lewontin RC (1979) The spandrels of san marco and the panglossian paradigm: a critique of the adaptationist programme. Proceedings of the Royal Society of London Series B Biological Sciences 205: 581–598.

24. Alvarez-Buylla E, Benítez M, Espinosa-Soto C, et al. (2007) Phenotypic evolution is restrained by complex developmental processes. HFSP journal 1: 99–103.

25. Gould SJ (1983) Hen's teeth and horse's toes. WW Norton & Company.

26. Hallgrímsson B, Hall BK (2011) Variation: a central concept in biology. Academic Press.

27. Reid RG (2007) Biological emergences: Evolution by natural experiment. MIT Press.

28. Shapiro JA (2011) Evolution: a view from the 21st century. Pearson Education.

29. Shapiro JA (2012) Rethinking the (im) possible in evolution. Progress in Biophysics and Molecular Biology .

30. Wilkins AS (2008) Waddington's unfinished critique of neo-darwinian genetics: Then and now. Biological Theory 3: 224–232.

31. Huang S (2011) Systems biology of stem cells: three useful perspectives to help overcome the paradigm of linear pathways. Philosophical Transactions of the Royal Society B: Biological Sciences 366: 2247–2259.

32. Koonin EV (2011) The logic of chance: the nature and origin of biological evolution. FT press.

33. Davila-Velderrain J, Servin-Marquez A, Alvarez-Buylla ER (2013) Molecular evolution constraints in the floral organ specification gene regulatory network module across 18 angiosperm genomes. Molecular biology and evolution : mst223.

34. Lewontin R (2011) The genotype/phenotype distinction. Stanford Encyclopedia of Philosophy.

35. Huang S, Kauffman S (2009) Complex gene regulatory networks-from structure to biological observables: cell fate determination. Encyclopedia of Complexity and Systems Science Meyers RA, editors Springer : 1180–1293.

36. Kaneko K (2011) Characterization of stem cells and cancer cells on the basis of gene expression profile stability, plasticity, and robustness. Bioessays 33: 403–413.

37. Mendoza L, Alvarez-Buylla ER (1998) Dynamics of the genetic regulatory network for¡ i¿ arabidopsis thaliana¡/i¿ flower morphogenesis. Journal of theoretical biology 193: 307–319.

38. Espinosa-Soto C, Padilla-Longoria P, Alvarez-Buylla ER (2004) A gene regulatory network model for cell-fate determination during arabidopsis thaliana flower development that is robust and recovers experimental gene expression profiles. The Plant Cell Online 16: 2923–2939.

39. Sanchez-Corrales YE, Alvarez-Buylla ER, Mendoza L (2010) The arabidopsis thaliana flower organ specification gene regulatory network determines a robust differentiation process. Journal of theoretical biology 264: 971–983.

40. Kauffman SA (1993) The origins of order: Self-organization and selection in evolution. Oxford university press.

41. Huang S, Eichler G, Bar-Yam Y, Ingber DE (2005) Cell fates as high-dimensional attractor states of a complex gene regulatory network. Physical review letters 94: 128701.

42. Cotterell J, Sharpe J (2013) Mechanistic explanations for restricted evolutionary paths that emerge from gene regulatory networks. PloS one 8: e61178.

43. Waddington CH (1957) The strategy of genes. London: George Allen & Unwin, Ltd.

44. Álvarez-Buylla ER, Chaos Á, Aldana M, Benítez M, Cortes-Poza Y, et al. (2008) Floral morphogenesis: stochastic explorations of a gene network epigenetic landscape. Plos one 3: e3626.

45. Zhou JX, Qiu X, dHerouel AF, Huang S (2014) Discrete gene network models for understanding multicellularity and cell reprogramming: From network structure to attractor landscapes landscape. In: Computational Systems Biology Second Edition Elsevier : 241–276.

46. Li C, Wang J (2013) Quantifying waddington landscapes and paths of non-adiabatic cell fate decisions for differentiation, reprogramming and transdifferentiation. Journal of The Royal Society Interface 10: 20130787.

47. Barrio RÁ, Hernandez-Machado A, Varea C, Romero-Arias JR, Alvarez-Buylla E (2010) Flower development as an interplay between dynamical physical fields and genetic networks. PloS one 5: e13523.

48. Barrio RA, Romero-Arias JR, Noguez MA, Azpeitia E, Ortiz-Gutiérrez E, et al. (2013) Cell patterns emerge from coupled chemical and physical fields with cell proliferation dynamics: the arabidopsis thaliana root as a study system. PLoS computational biology 9: e1003026.

Chapter 3

Metodología

A Description of Phenomena is Not Equivalent to an Understanding

... an understanding of some phenomenon is not obtained by constructing and adjusting a set of equations in such a manner that it provides an accurate model.

It is much more meaningful scientifically to seek the construction of simple models and endeavor to derive an understanding from these than to attempt to mimic every detail of each specific system we encounter
— Kunihiko Kaneko, *Life: An Introduction to Complex Systems Biology* (2006)

Chapter 26

Gene Regulatory Network Models for Floral Organ Determination

Eugenio Azpeitia, José Davila-Velderrain, Carlos Villarreal, and Elena R. Alvarez-Buylla

Abstract

Understanding how genotypes map unto phenotypes implies an integrative understanding of the processes regulating cell differentiation and morphogenesis, which comprise development. Such a task requires the use of theoretical and computational approaches to integrate and follow the concerted action of multiple genetic and nongenetic components that hold highly nonlinear interactions. Gene regulatory network (GRN) models have been proposed to approach such task. GRN models have become very useful to understand how such types of interactions restrict the multi-gene expression patterns that characterize different cell-fates. More recently, such temporal single-cell models have been extended to recover the temporal and spatial components of morphogenesis. Since the complete genomic GRN is still unknown and intractable for any organism, and some clear developmental modules have been identified, we focus here on the analysis of well-curated and experimentally grounded small GRN modules. One of the first experimentally grounded GRN that was proposed and validated corresponds to the regulatory module involved in floral organ determination. In this chapter we use this GRN as an example of the methodologies involved in: (1) formalizing and integrating molecular genetic data into the logical functions (Boolean functions) that rule gene interactions and dynamics in a Boolean GRN; (2) the algorithms and computational approaches used to recover the steady-states that correspond to each cell type, as well as the set of initial GRN configurations that lead to each one of such states (i.e., basins of attraction); (3) the approaches used to validate a GRN model using wild type and mutant or overexpression data, or to test the robustness of the GRN being proposed; (4) some of the methods that have been used to incorporate random fluctuations in the GRN Boolean functions and enable stochastic GRN models to address the temporal sequence with which gene configurations and cell fates are attained; (5) the methodologies used to approximate discrete Boolean GRN to continuous systems and their use in further dynamic analyses. The methodologies explained for the GRN of floral organ determination developed here in detail can be applied to any other functional developmental module.

Key words Gene regulatory networks, Functional module, Flower development, Cell differentiation, Attractors, Morphogenesis, Dynamics, Floral organ determination, Attractors, Basins of attraction, Stochastic networks, Mathematical models, Computational simulations, Robustness

---

Eugenio Azpeitia, José Davila-Velderrain, and Carlos Villarreal contributed equally to this work.

José Luis Riechmann and Frank Wellmer (eds.), *Flower Development: Methods and Protocols*, Methods in Molecular Biology, vol. 1110, DOI 10.1007/978-1-4614-9408-9_26,

1 Introduction

The mapping of the genotype unto the phenotypes implies the concerted action of multiple components during cell differentiation and morphogenesis that comprise development [1]. These components are part of regulatory motifs, which hold nonlinear interactions that produce complex behaviors [2, 3]. Such complexity cannot be understood in terms of individual components, and rather emerges as a result of the interactions among the components of the whole system. In order to integrate the action of multiple molecular components and follow their dynamics, it is indispensable to postulate mathematical and computational models. Gene regulatory network (GRN) models have appeared as one of the most powerful tools for the study of complex molecular systems. Small GRNs can sometimes be studied with analytical mathematical formulations, while medium or large size GRNs are amenable for dynamical analyses only with computer simulations [4]. As following the dynamics of the genomic interactomes is still intractable even with the most powerful computers, and given the fact that genomic networks are composed of multiple structural and functional modules, others and we have proposed to search for such modules for the study of biomolecular systems dynamics using GRN models (e.g., [5–7]).

Boolean models are probably the simplest type of formalism employed for the study of GRNs. Nonetheless, Boolean models provide meaningful information about the system. Importantly, Boolean GRNs can be approximated to continuous models that enable the use of additional mathematical tools [4, 8]. Given that: (a) the logic of GRNs is adequately formalized with Boolean models; (b) obtaining real biological parameters from biological molecular systems is still a complicated task; and (c) the use of realistic models can be computationally expensive, we believe that Boolean models and their continuous approximations are becoming a fundamental and practical tool to study GRN dynamics and to understand the complex behaviors observed in developmental processes (*see* refs. 9–11).

Based on the above rationale, the first step in building a GRN model is the identification of a developmental module and the integration of all the experimental data on the molecular components participating in it. The ABC genetic model of floral organ determination (*see* refs. 3, 12) (*see* Chapter 1) is part of a clearly circumscribed developmental module that underlies the subdifferentiation of the floral meristem in four concentric rings early on during flower development. From the outermost part of the floral meristem to its center, each ring comprises the primordial cells of sepals, petals, stamens, and carpels. Based on experimental

evidence [13], it became obvious that although necessary, the ABC genes are not sufficient to specify floral organs. The ABC model has been instrumental to understanding flower development and evolution. However, it does not constitute a dynamic model able to recover the ABC combinatory code, as well as explain how the expression profiles of the set of molecular components included in the flower organ determination GRN, which includes the ABC genes, is established to promote the sepal, petal, stamen and carpel cell fates. Importantly, such a dynamic GRN model is the basis to understand how such cell types are determined in time and space, and thus, how the morphogenetic pattern that characterizes young floral meristems will form adult flowers [12, 14].

In order to uncover the necessary and sufficient set of interacting components involved in floral organ specification, the first step implies recovering the experimental evidence of ABC gene interacting components that include both regulated and regulator genes. In the case of Boolean models, the experimental data is formalized in the form of Boolean functions, which determine the dynamics of the GRN. In Boolean or any other type of discrete network, it is possible to fully explore the whole set of configurations or states of the system, and find the steady state configurations (attractors; see below). Kauffman postulated that the attractors to which GRNs converge, could correspond to the states characterizing differentiated cells [15]. More recently, Boolean GRNs have been grounded on experimental data ([5]; *see* review in ref. 3) showing that the attractors of developmental networks indeed correspond to the stable gene configuration observed in different types of cells, as long as a sufficient set of components involved in a given developmental module are incorporated.

In this Chapter we focus on the regulatory module underlying floral organ determination in *Arabidopsis thaliana* during early stages of flower development. Some of the methodologies explained here have been used in previous publications on such GRN [5, 7, 16–19]. In this chapter we will use examples extracted mainly from our own studies to explain how to develop and extend experimentally supported Boolean GRN models. Then, we explain how to incorporate stochastic properties in the model, which can allow us to explore the temporal sequence with which attractors or cell gene configurations and cell-fates are attained (e.g., [4]). Finally, we explain how we can approximate the Boolean model to a continuous one that can then be used in other types of models, for example, to explore spatial aspects of morphogenesis [14]. It is important to keep in mind that the tools presented in this Chapter can be applied to any GRN. Consequently, we begin with general explanations and afterwards we use examples from the literature to illustrate each methodological step.

2 Methods

2.1 Definitions

GRN nodes and edges: In GRNs, nodes represent genes, proteins or other types of molecular components such as miRNAs and hormones, while edges represent regulatory interactions among the components. Usually the interactions are positive (activations) or negative (inhibitions), but other type of interactions can be included (e.g., protein-protein interactions).

Variables: Variables are the elements that describe the system under study (usually the nodes) and which can take different values at each time.

Variable/Gene state: The value that a node takes at a certain time represents its state. The state can be a discrete or continuous value. In the case of Boolean networks the states can only be "0" when "OFF" and "1" when "ON."

Network State/Configuration: The vector composed by a set of values, where each value corresponds to the state of a specific gene of the network. In a Boolean network such vectors or network configurations are arrays of "0's" and "1's."

Attractors: Stationary network configurations are known as attractors. Single-state, stationary configurations are known as fixed-point attractors (Fig. 1a) and these are generally the ones that correspond to the arrays of gene activation states that characterize

a

Time	GEN1	GEN2	GEN3
1	1	0	0
2	1	0	0
3	1	0	0
.			
.			
.			
n-1	1	0	0
n	1	0	0

Fixed-point attractor

b

Time	GEN1	GEN2	GEN3
1	0	1	1
2	1	0	1
3	0	1	1
.			
.			
.			
n-1	1	0	1
n	0	1	1

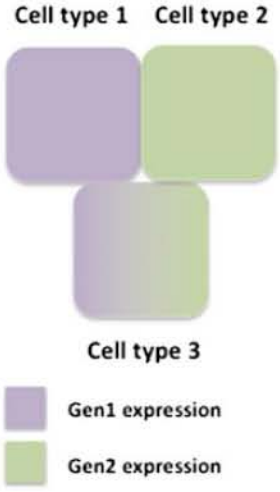
Cyclic attractor

c

Time	GEN1	GEN2	GEN3
1	1	1	1
2	1	1	0
3	0	0	1
.			
.			
.			
n-1	0	0	0
n	0	1	0

Transitory states

Fig. 1 Fixed-point attractors, cyclic attractors, and transitory states. (**a**) An example of a fixed-point attractor. As observed, fixed-point attractors have one unique state where they stay indefinitely unless something perturbs them. (**b**) An example of a cyclic attractor. Cyclic attractors are composed of two or more network states that orderly repeat. In this case we observe a two state cyclic attractor. (**c**) Transitory states. Transitory states are states that lead to an attractor, but are not attractors themselves



Expected attractors	GEN1	GEN2
Cell type 1	1	0
Cell type 2	0	1
Cell type 3	1	1

Fig. 2 The set of expected attractors. As explained in the main text, the set of expected attractors is obtained from the experimental information. In the case of cell types, the attractors correspond to the observed stable gene configuration of each cell type. Thus, if our system consists in three different cell types, one cell type with GEN1 expression, other with GEN2 expression, and a third one with both GEN1 and GEN2 expression, our set of expected attractors will be exactly this

different cell types. Whereas a set of network states that orderly repeat cyclically correspond to cyclic attractors (Fig. 1b).

Transitory states: All states that are not or do not form part of an attractor are transient or transitory states (Fig. 1c).

Basin of attraction: The set of all the initial configurations that eventually lead to a particular attractor constitute its basin of attraction.

Expected or observed attractors: Gene expression profiles or configurations that have been obtained from experimental assays and reported in the scientific literature for particular cell types are referred to here as the expected or observed attractors. Such attractors are expected to be recovered by the postulated GRN (Fig. 2).

Model Validation: The task of evaluating a model by means of contrasting its predictions with experimental results. For Boolean GRNs, model validation would imply, among others: recovering the observed gene configurations for the cells under study under *wt* and mutant or overexpression conditions, robustness analyses, etc. (see below).

Robustness: The ability of a system to maintain an output in the face of perturbations. For the case of a Boolean GRN model, it is evaluated, for example, by assessing if the system's attractors are still recovered under different transient and permanent mutations (alterations in the Boolean functions, nodes, or GRN topology).

2.2 General Protocol

A generic protocol to postulate a GRN model for a particular developmental module would be as follows:

(i) Identify a structural or functional developmental module (*see* **Note 1**).

(ii) Based on available experimental data, select the set of potential nodes or molecular components that will be incorporated in the GRN model with the aim of integrating the key necessary and sufficient components of the functional module under analysis. Then, explore the experimental data concerning the spatio-temporal expression patterns of the genes to be incorporated in the model and assemble a table with a Boolean format of the expected configurations that should be recovered with the GRN model (such configurations are the "expected attractors") (*see* **Note 2**).

(iii) Integrate and formalize the experimental data concerning the interactions among the selected nodes using Boolean logical functions that will rule the Boolean GRN dynamics.

(iv) The GRN is modeled as a dynamic system by exploring the states attained, given all possible initial configurations and the Boolean functions defined in (iii). The GRN is initialized in all possible configurations and followed until it reaches a fixed-point or cyclic attractor (*see* **Note 3**).

(v) Compare the simulated attractors to the ones observed experimentally (expected attractors; *see* **item** (**ii**) above). A perfect coincidence would suggest that a sufficient set of molecular components (nodes) and a fairly correct set of interactions have been considered in the postulated GRN model. If this is not the case, additional components and interactions can be incorporated or postulated, or the Boolean functions can be modified. This allows to refine interpretations of experimental data, or to postulate novel interactions to be tested experimentally in the future. In any case, the process can be repeated several times based on the dynamical behavior of the modified versions of the GRN under study until a regulatory module is postulated. Such module can include some novel hypothetical interactions or components, integrate available experimental data, and identify possible experimental contradictions or holes.

(vi) To validate the model, it is addressed if it recovers the *wt* and mutant (loss of function and gain of function) gene activation configurations that characterize the cells being considered. Perturbation analyses of the nodes and interactions, or the Boolean functions, can also be used for validating the model in order to test the robustness of the GRN under study. Eventually, novel predictions can be made and tested experimentally.

(vii) To recover the dynamics of the GRN and the temporal pattern of attractor attainment, the logical functions can be modeled as stochastic ones. Observed temporal patterns of cell-fate or gene configurations attainment can be used to validate the GRN model under consideration.

(viii) For further applications and also in cases that continuous functions are appropriate to describe the behavior of some of the components, the Boolean model can be approximated to a continuous one (*see* Subheading 2.5). Besides being useful for further modeling procedures, the continuous approximation is also a means of performing a robustness analysis of the GRN under study. Such a task hence implies as well a further validation of the model being postulated.

(ix) Equivalent approaches to the ones summarized in (vi) and (vii) for discrete systems can be used in continuous ones.

There are two types of materials needed when modeling dynamic GRNs. First, the expected results to be recovered by the model that are extracted from the literature and depend on the aims of the model and the nature of the developmental module being considered, but generally include stable gene configurations (attractors), mutant phenotypes, and developmental transitions, to name a few. The second set is the software required for the analyses of the GRN. Currently there are several available programs for GRN analyses (*see* **Note 8**). In the following sections, we explain with more detail and specific examples how this general protocol can be applied. We start by explaining the simplest Boolean approach for dynamical GRN modeling.

2.3 Deterministic Boolean GRN Model

In Boolean GRN models, nodes can only attain one of two possible values: "1" if the node is "ON," and "0" if the node is "OFF." A "0" node value usually represents that a gene is not being expressed, but can also represent the absence of a protein or hormone, while a "1" node value represents that a gene is expressed or another type of molecular component is present. As mentioned above, the first step in building a network is to extract the necessary experimental information to define the set of components to be considered in the GRN model, the set of expected attractors, and the Boolean functions that formally integrate the experimental data and define the dynamics of the GRN.

2.3.1 Expected Attractors

In Boolean GRNs, the network states (*see* Subheading 2.1) are defined by vectors of 0s and 1s. While a formal mathematical definition of attractors can be found on the chapter "Implicit Methods for Qualitative Modeling of Gene Regulatory Networks" of another Springer Protocols book [20], in Subheading 2.1 we give a more pragmatic definition of attractors, and we prefer to stick to it. In 1969, Kauffman proposed that the attractors of a GRN model

could correspond to stable gene configurations characteristic of particular cell types or physiological states (*see* Subheading 2.1; Fig. 2). Consequently, the expected attractors are defined from gene expression patterns obtained from the literature, as well as from other data sources that clearly define the spatio-temporal gene configuration of the system. For example, Espinosa-Soto and collaborators [7] defined the expected attractors from the gene expression patterns reported in scientific publications. In another study, La Rota and collaborators [19] integrated experimental data into a gene expression map for the sepal primordium. Based on its expression map they defined zones with different combinations of gene expression, and each zone corresponded to an expected attractor. Defining the expected set of attractors is an indispensable step when building the GRN model, because they are used to validate the GRN (see below). Although it should be clear that the postulation of the Boolean functions is an independent task, and hence, it does not imply circularity.

2.3.2 Boolean Functions

In a Boolean GRN model the state of expression of each gene changes along time according to the dynamic equation

$$x_i(t+\tau) = f_i\left(x_1(t), x_2(t), \ldots, x_k(t)\right), \quad (1)$$

in which the future state of gene i evolves temporally as a function of the current state of its k regulators. Boolean functions f_i can be formalized as logical statements or as truth tables. Logical statements use the logical operators "AND," "OR" and "NOT" to describe gene interactions, while in truth tables the state of the gene of interest is given for all possible state combinations of its k regulators (*see* **Note 4**). Logical operators can be combined in order to describe complex gene regulatory interactions, and can always be translated into an equivalent truth table. In Fig. 3, we provide examples of common gene regulatory interactions formalized as logical statements with their equivalent truth table. Consequently, in general, Boolean functions are generated from experimental evidence (but *see* **Note 5**). For example, if TGEN (a target gene) is ectopically expressed in a GEN1 loss-of-function background, it is inferred that GEN1 is a negative regulator of TGEN, and we use the "NOT" logical operator to describe GEN1 regulation over TGEN or its equivalent truth table (Fig. 4). In this Boolean function, the state of TGEN at time $t+\tau$ is 1 if GEN1 value is 0 at time t, and TGEN value at time $t+\tau$ is 0 if GEN1 value is 1 at time t (*see* **Note 6**).

The Boolean functions of the GRN developmental module being used here as an example, were grounded on available experimental information [5, 7, 17–19]. As with expected attractors, Boolean functions can be grounded on different types of experimental data, as long as they clearly state how genes interact (*see* **Note 7**). We now will provide an example of how the

GEN1(t)	GEN2(t)	TGEN(t+τ)
0	0	0
0	1	0
1	0	0
1	1	1

TGEN(t+τ) = GEN1(t) & GEN2(t)

GEN1(t)	GEN2(t)	TGEN(t+τ)
0	0	0
0	1	1
1	0	1
1	1	1

TGEN(t+τ) = GEN1(t) | GEN2(t)

GEN1(t)	GEN2(t)	TGEN(t+τ)
0	0	0
0	1	0
1	0	1
1	1	0

TGEN(t+τ) = GEN1(t) & ! GEN2(t)

GEN1(t)	GEN2(t)	TGEN(t+τ)
0	0	0
0	1	1
1	0	0
1	1	0

TGEN(t+τ) = ! GEN1(t) & GEN2(t)

Fig. 3 Examples of common Boolean functions. Here we present four examples of common Boolean functions for a target gene, in this case TGEN, with two regulators, namely, GEN1 and GEN2

experimental information was integrated and formalized as a Boolean function. During the transition from inflorescence to flower meristem, the expression of *TERMINAL FLOWER 1* (*TFL1*) needs to be repressed [21, 22], because *TFL1* is a promoter of inflorescence development [23]. *TFL1* is transcribed in the center of the meristem and from there it moves to peripheral cells [24]. *EMF1* is assumed to be a positive regulator of *TFL1* because the *emf1* mutant is epistatic to *tfl1* loss-of-function mutant, and both, *tfl1* and *emf1* mutants have similar phenotypes in terms of inflorescence meristem identity [25]. The over expression phenotype of *AP1* is similar to the loss-of-function of *TFL1*, and in the *ap1* mutant *TFL1* is ectopically expressed, suggesting that *AP1* is a negative regulator of *TFL1* [26]. Similarly, *TFL1* expression is not observed in *LFY* over expression and is ectopically expressed in *LFY* loss-of-function mutants [27]. According to these results, *EMF1* is a positive regulator of *TFL1*, while *AP1* and *LFY* are

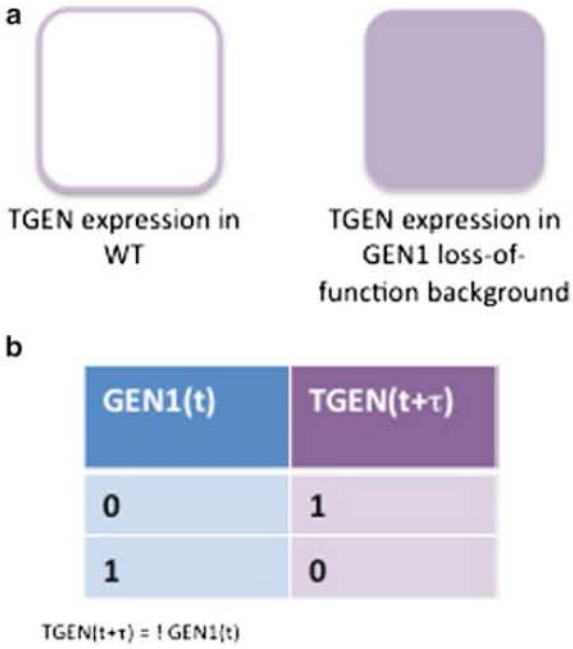


Fig. 4 Truth table and logical statement of the example explained in the main text. (**a**) TGEN expression is not observed in the GEN1 loss-of-function background. Hence, we can assume that GEN1 is a negative regulator of TGEN. This Boolean function can be represented with a (**b**) truth tab. or a (**c**) logical statement

negative regulators of *TFL1*. These results were formalized as a logical statement [18] as follows:

$$TFL1 = EMF1\ AND\ NOT\ AP1\ AND\ NOT\ LFY$$

A complete list of the Boolean functions and the experimental evidence for this model can be found in refs. 7, 18; *note some typographical errors corrected in* refs. 1, 12.

2.3.3 Validating the GRN: Simulated Attractors vs. Expected Attractors

Once the Boolean functions and the set of expected attractors of the GRN are obtained, we can proceed to make a first, necessary validation of the GRN. The first step is to use numerical simulations to recover the attractors that our set of Boolean functions generates (*see* **Note 8**). The attractors recovered in the simulations must coincide with the expected attractors, based on experimental data. In Espinosa-Soto and collaborators [7] ten attractors were recovered. Four out of the ten attractors corresponded to gene activation configurations that characterize meristematic cells of inflorescence meristems, while the rest corresponded to the gene configurations observed in sepal, petal, stamen and carpel primordial cells (Fig. 5). In the GRN for sepal development formulated by La Rota and collaborators [19], at least two attractors were recovered; one corresponding to the abaxial and the other one to the adaxial cells of the floral organ.

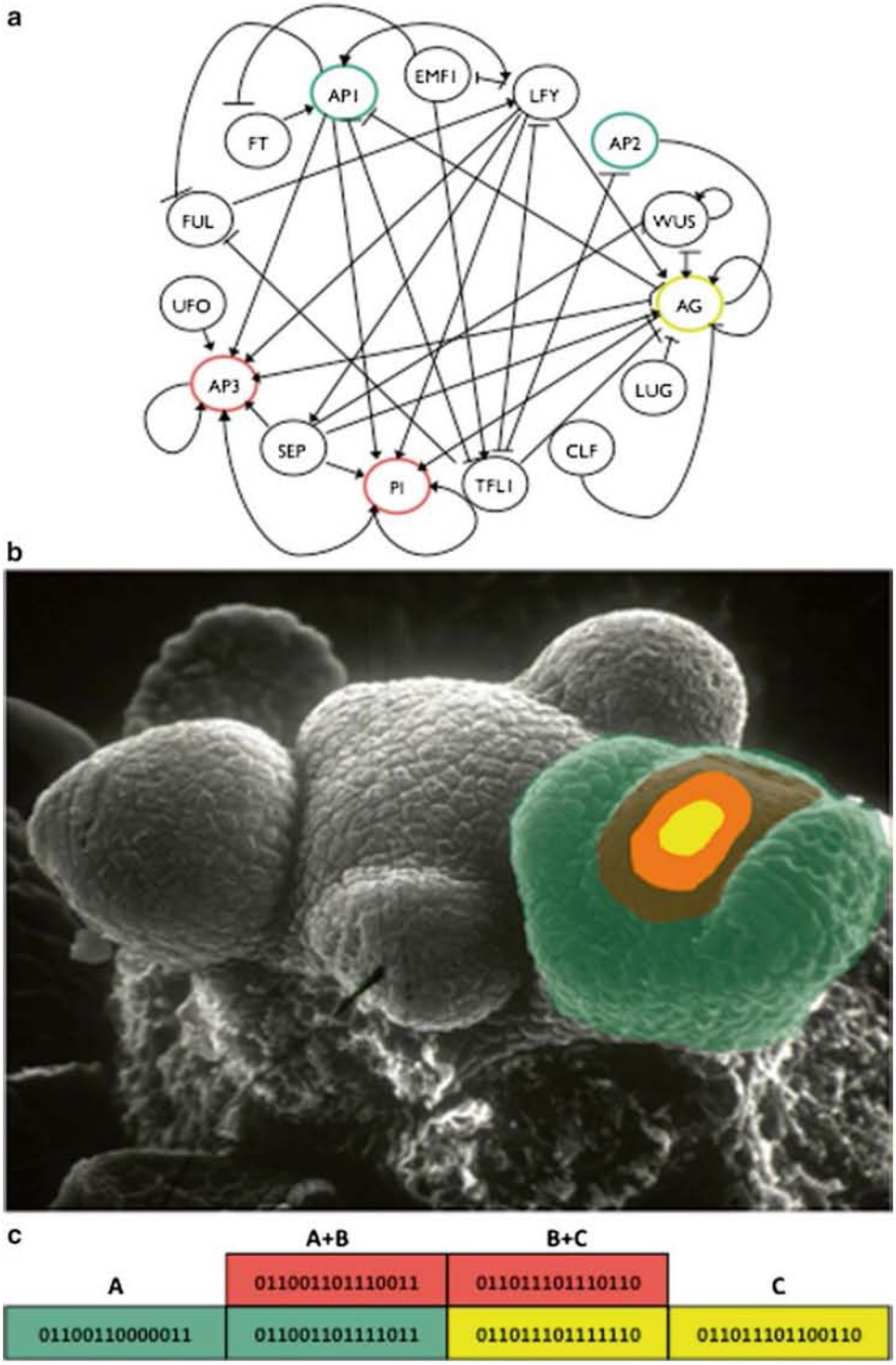


Fig. 5 Obtained attractors of the flower organ specification GRN. In (**a**) we present the graph of the flower organ specification GRN proposed by Espinosa-Soto and collaborators [7]. The GRN recovered 10 fixed-point attractors. Six of the attractors corresponded to the observed gene configuration in the primordial cell of sepals (one attractor), petals (two attractors), statements (two attractors), and carpels (one attractor). (**b**) A flower meristem in which the primordial sepal cells are colored in *green*, primordial petal cells in *brown grey*, primordial stamens in *orange*, and primordial carpel cells in *yellow*. In (**c**), the ABC model and the floral organ determination GRN attractors that correspond to A, A + B, B + C, and C gene combinations, which specify sepal, petal, stamen, and carpel primordial cells, respectively. The activation states correspond to each of the GRN nodes starting on the left with "EMF1" and consecutively progressing clockwise the rest of the genes in the GRN shown in (**a**)

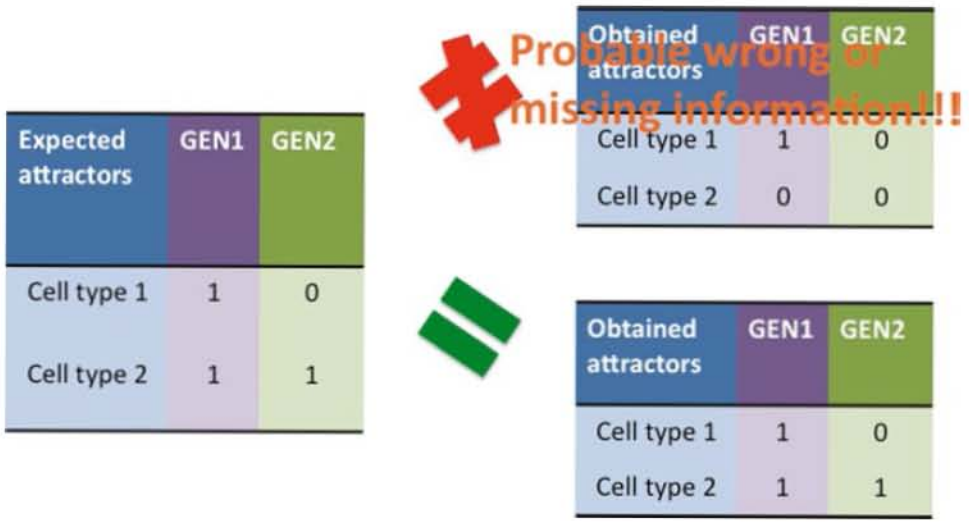


Fig. 6 The set of expected attractors vs. the set of obtained attractors. Both the set of expected and obtained attractors must coincide, when this do not happens it is usually assumed that there is some wrong or missing information

In cases in which the attractors recovered by the simulated GRN under study and those observed experimentally do not coincide, additional nodes or interactions can be considered, or the postulated Boolean functions can be modified (Fig. 6). Such novel hypotheses can be tested by running the GRN dynamics once more, and if the simulated and observed (expected) attractors now coincide, the model can be used to postulate novel interactions, missing data, or contradictions among those that had been proposed previously. For example, in Espinosa-Soto and collaborators [7] four missing interactions were predicted. Importantly, some of these predictions have been experimentally validated by independent and posterior research, demonstrating the predictive capacity and usefulness of this approach.

2.3.4 *Mutant Analysis*

An additional means to validate a GRN model is to simulate loss-of-function (fixing the mutated gene expression value to 0) and gain-of-function (fixing the overexpressed gene expression value to 1) mutants. The recovered attractors in the model with such altered fixed expression values must correspond to the effects experimentally observed in the corresponding mutants (*see* Fig. 7; **Note 9**). If a discrepancy is found in such a validation process, additional hypotheses concerning new nodes or interactions can be postulated. For the postulated GRN module underlying floral organ determination, most of the recovered attractors in the simulated mutants corresponded to the genetic configurations that have been observed experimentally [7, 17, 18]. In some cases, the simulated and observed (expected) attractors did not coincide and new interactions were postulated. For example, in Espinosa-Soto and collaborators [7] a positive feedback loop was predicted for the

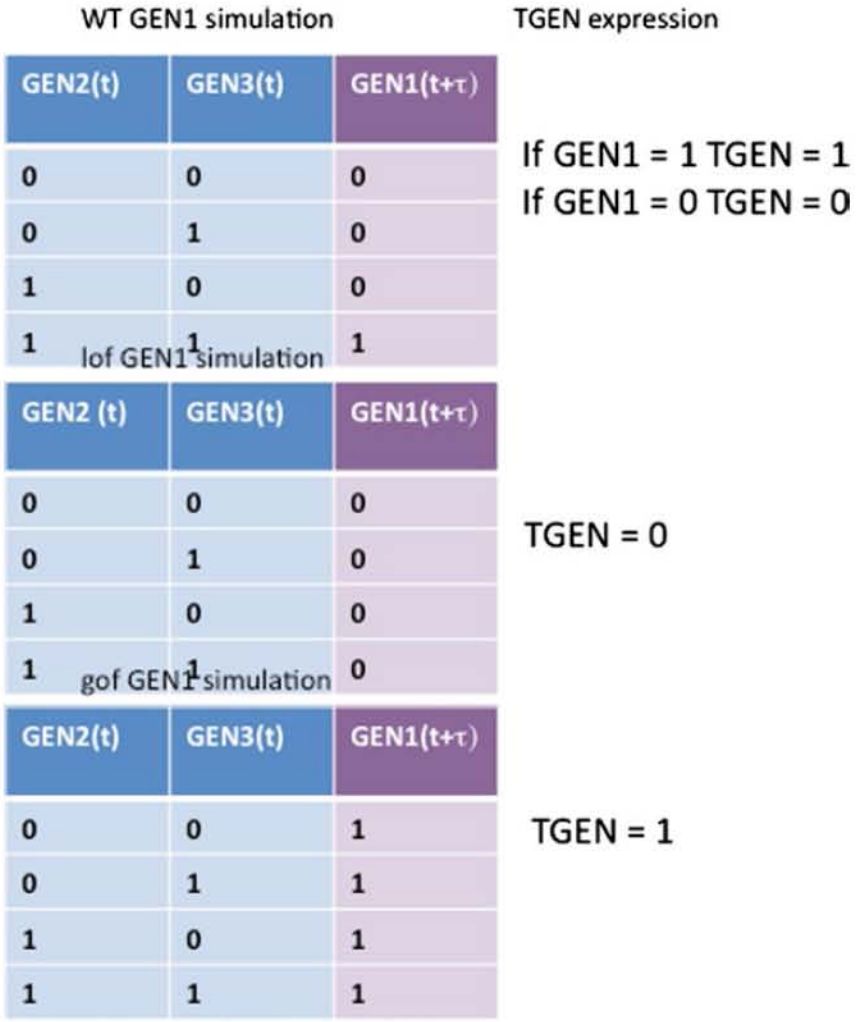


Fig. 7 Loss-of-function and gain of function mutant simulations. Loss-of-function and gain-of-function mutant simulations are done by fixing the state of the desired gene to 0 and 1, respectively. In (**a**) the Boolean function of a non-mutated GEN1. In (**b**) and (**c**) the Boolean function of the same gene in a loss-of-function and a gain-of-function simulation, respectively. The Boolean functions are presented as truth tables and as logical statements. lof = loss-of-function, gof = gain-of-function

gene *AGAMOUS* (*AG*), even though this seemed unlikely because in the *ag-1* loss-of-function mutant plants, the *AG* expression pattern is the same as in wild-type plants [28]. In a posterior study in an independent laboratory, the prediction was verified experimentally [29].

Simulations of mutants are also useful when trying to predict the effects of multiple mutants, which are complicated to generate in the laboratory. Moreover, even when the GRN involved in flower determination in *Arabidopsis* and *Petunia* seems to be conserved, the mutant phenotypes are not identical. Espinosa-Soto and collaborators [7] used mutant analyses to test the effect of a

duplication in B genes that has been reported in *Petunia*, and recovered the single mutant that had been described, and at the same time predicted the expected phenotype for the double mutant of the two duplicates.

2.3.5 Robustness Analyses

Experimental and theoretical work has demonstrated that living organisms are robust against perturbations. Moreover, at the molecular level the processes involved in different biological behaviors are also robust against internal and external variations. Such robustness implies that the overall functionality of the system remains when perturbed [30, 31]. In the case of GRNs, attractors should be robust when the Boolean functions are altered. In Espinosa-Soto and collaborators [7] the output value of every line of the truth tables was changed one by one. Interestingly, we found that the original attractors did not change for more than 95 % of the logical table alterations, indicating that the functionality of the postulated developmental module is robust to this type of perturbation. There are other types of perturbation analyses. For example, we could change with a certain probability the value of a line of the truth table, or the state of the network. Similarly, if we perturb the GRN with these other types of perturbations, the systems' attractors are expected to be maintained.

2.4 Stochastic Boolean GRN Model: Temporal Sequence of Cell-Fate Attainment

In deterministic GRN models, as the Boolean model exposed above, the system under study always converges to a single attractor if initialized from the same configuration, and once it attains such steady-state, it remains there indefinitely. However, during a developmental process, cells change from one stable cell configuration to another one in particular temporal and spatial or morphogenetic patterns. In order to explore questions such as how differentiating cells decide between one of the available attractors, or the order in which the system converges to the different attractors, given an initial condition, and to make statistical predictions of such possible behaviors, a stochastic formalism is needed.

In this section we develop a discrete stochastic model as an extension of the deterministic Boolean GRN. We then show how this approach can be used to explore the patterns of cell-fate attainment. Specifically, the model formalism explained here allows the investigation of the temporal sequence with which attractors are visited in the GRN when noise or random perturbations drive the system from one attractor to any other one.

2.4.1 From Deterministic to Stochastic Models

In a Boolean GRN model the dynamics given by Eq. 1 is deterministic: for a given set of Boolean functions f_i (*see* Subheading 2.3.2), the configuration of the network at time t completely determines the configuration of the network at the next time step $t+1$ (conventionally $\tau = 1$). If Eq. 1 is iterated starting from a given initial configuration (defined by an array of n entries with 0s and 1s

representing the activation states of the n genes), the network will eventually converge to an attractor. This deterministic version implies that once the system reaches an attractor, it remains there for all subsequent iterations. However, if noise is introduced into either the Boolean functions, or the gene states, there is a finite probability for the system to "jump" from one basin of attraction to another one (for definitions, *see* Subheading 2.1) and consequently, from one attractor to another one. Such a stochastic Boolean model of the GRN enables the study of transitions among attractors.

Noise can be implemented in a Boolean GRN model in several ways (*see* **Note 10**). Here we implement noise by introducing a constant probability of error ξ for the deterministic Boolean functions. In other words, at each time step, each gene "disobeys" its Boolean function with probability ξ, such that in the stochastic version, Eq. 1 is extended to

$$x_i(t+\tau) = \begin{cases} f_i(t), & \textit{with prob.} 1-\xi \\ 1-f_i(t) & \textit{with prob.} \xi \end{cases} \quad (2)$$

Note that the stochastic version (e.g., Eq. 2) reduces to a deterministic one (Eq. 1)) when $\xi = 0$. In the model, the stochastic perturbations are applied independently and individually to each gene at each iteration. This implementation of noise for stochastic Boolean modeling of GRNs has been referred to as the stochasticity in nodes (SIN) model with the assumption of a single fault at a time [20, 32].

2.4.2 The Transition Probability Matrix

When Eq. 2 is iterated, both the set of Boolean functions f_i and the error probability ξ determine the configuration of the network at the next time step. Under this stochastic dynamics, a given initial configuration will no longer converge to the same attractor each time. This situation allows us to estimate a probability of transition from one network state to another state as the frequency with which this transition occurs in a large number of repetitions of the same iteration (see below). The estimated transition probabilities can then be used to study the behavior of the system and to make statistical predictions.

As we want the model to be useful in the exploration of the patterns of temporal cell-fate attainment, the network states that we are interested in are the fixed-point attractor states that represent the cell types. Thus, we need to estimate the probability p_{ij} of transition from the attractor i to the attractor j. From the deterministic Boolean model, we already know to which attractors the network converges. In the following we use the term attractor to refer to both, the attractor and its basin. Thus, we can define a scalar (single-valued) variable X_t to describe the state of the network in terms of the specific attractor in which the network is in at

time t. Then, X_t will take at time t any value from the ordered set $(1,2,i,\ldots,K)$ where each i represents one specific attractor from the available k attractors. The configuration of the network at time t is then related to the configuration at time $t+1$ through what is known as the transition probabilities. If the network is in attractor i at time t, at the next time step $t+1$, it will either stay in attractor i or move to another attractor j.

Formally, p_{ij} denotes a one-step transition probability that is defined as the following conditional probability:

$$p_{ij} = Prob\{X_{t+1} = j \mid X_t = i\}, \tag{3}$$

the probability that the network at time $t+1$ is in the attractor j given that it was in the attractor i at the previous time t, where $i,j=1,2,\ldots,K$ for K attractors. The set of probabilities p_{ij} can be expressed in matrix form:

$$P = \begin{pmatrix} p_{11} & \cdots & p_{1k} \\ \vdots & \ddots & \vdots \\ p_{k1} & \cdots & p_{kk} \end{pmatrix}.$$

As the number of attractors K is finite, P is a $K \times K$ transition matrix. Operationally, under the current model, one can estimate the probabilities of the i-th row by first iterating Eq. 2 one time step starting from a given initial configuration corresponding to the basin of attraction of attractor i. If, after the iteration, the system remains in the same attractor, or the same basin of attraction, one count is added to the diagonal entry that corresponds to P_{ii}. If the configuration ends up in a different basin j, the count is added to the column j that corresponds to p_{ij}. This process is repeated a large number of times (e.g., 10,000) for each of the possible $\Omega = 2^n$ initial conditions. For each state (attractor), the one-step transition probabilities should satisfy $\sum_{j=i}^{K} p_{ij} = 1$ and $p_{ij} \geq 0$. This means that in the transition matrix P, the rows must sum to 1. This is achieved by dividing the number of counts in each matrix entry by the total number of configurations that started in the corresponding matrix row (e.g., basin i). As the dynamics in Eq. 2 are driven by both the Boolean functions f_i and the error probability ξ, given a fixed set of Boolean functions, different values of ξ will result in different values of the transition probabilities p_{ij} (*see* **Note 11**).

2.4.3 The Probabilistic Dynamics of Cell-Fate Attainment

Once the transition matrix P is calculated, it can be used in a dynamic model to describe how the probability of being in a particular attractor changes in time. In other words, we are now in position to derive a probabilistic dynamic model to simulate the dynamics of temporal cell-fate attainment.

In the previous subsection, the dynamics of transition between attractor states were defined in terms of transition probabilities.

When this is the case, the state of the network at any given time X_t can only be represented by its associated discrete probability distribution. We denote this distribution by the vector $p_X(t)=(p_1(t), p_2(t),\ldots,p_K(t))$, where $p_i(t)$ represents the probability of the network being in attractor i at time t, and $\sum_{i=i}^{k} p_i(t)=1$.

Given $p_X(t)$, the probability distribution associated with X_{t+1} can be found by multiplying the transition matrix P by $p_X(t)$. We obtain the following dynamic equation

$$p_x(t+1)=p_x(t)P, \tag{4}$$

this latter equation projects the process forward in time, and it allows us to follow the dynamics of the probabilities of cell-fate attainment by means of straightforward iteration.

In order to do so, it is necessary to specify an initial vector $p_x(t=0)$ which represents the probability distribution of the network state at time $t=0$. In biological terms, this initial vector can be interpreted as the representation of how a large population of cells is distributed over the available attractors. In other words, how many cells of each type are in the population at the initial time $t=0$. As the probabilities p_i sum to one, an underlying assumption is that the number of cells in the population remains constant. In the next subsection we show how this initial distribution can be chosen based on a biological motivation in order to explore a specific question regarding the dynamics of cell-fate attainment during floral organ formation. When the matrix P and the initial vector $p_X(0)$ are specified, Eq. 4 can be iterated (*see* **Note 12**); this process will generate a trajectory for the temporal evolution of the probability of each of the attractors. Every attractor will have a maximum in the probability of being reached at particular times. This maximum corresponds to the moment at which the corresponding cell-fate is most likely. Thus, the order in which the maximal probability of the different attractors is reached may serve as an intrinsic explanation for the emerging temporal order during early stages of development. Note that, as the transition probabilities of the matrix P depend on the value of ξ used in Eq. 2, the trajectories for the probability of attractor attainment will vary for different values of the error probability ξ.

2.4.4 Temporal Cell-Fate Pattern During Early Stages of Flower Development

In this subsection we show how the modeling formalism presented above can be applied to propose mechanistic explanations of observed patterns of temporal cell-fate attainment. In the modeling framework presented here, stochasticity may seem just as a modeling artifact that allows the study of transitions among attractors. However, a multitude of studies have demonstrated both theoretically and experimentally that stochasticity and the so-called biological noise are ubiquitously present in biological systems given the chemical nature of biological processes (for example *see* refs. 33–36).

Table 1
Example of a transition matrix *P* estimated from the GRN model for the floral organ determination of *A. thaliana*. The matrix elements are the transition probabilities among pairs of the six attractors (S, P1, P2, S1, S2, and C). Probabilities where calculated in Alvarez-Buylla et al. [4] using ($\xi = 0.01$)

	sep	pe1	pe2	st1	st2	car
sep	0.939395	0.001943	0.009571	0.000083	0.000490	0.048517
pe1	0.036925	0.904162	0.009250	0.033900	0.000488	0.015275
pe2	0.009067	0.000464	0.941609	0.000024	0.048374	0.000461
st1	0.000084	0.001893	0.000020	0.936514	0.009960	0.051530
st2	0.000020	0.000001	0.002074	0.000356	0.987953	0.009597
car	0.002045	0.000034	0.000020	0.001951	0.010020	0.985930

Under the hypothesis that random fluctuations in a system may be important for cell behavior and pattern formation, Alvarez-Buylla and collaborators proposed a discrete stochastic model to address whether noisy perturbations of the GRN model for the floral organ determination of *A. thaliana* are sufficient to recover the stereotypical temporal pattern in gene expression during flower development [4]. As mentioned above, previous analysis of the deterministic Boolean GRN showed that the system converges only to ten fixed-point attractors, which correspond to the main cell types observed during early flower development [7]. Six of the attractors correspond to the four floral organ primordial cells within the flower meristem: sepals, petals, stamens, and carpels (S, P1, P2, S1, S2, and C).

Following Subheading 2.4.2, we can study the dynamics of cell-fate attainment of the floral organ primordial cells by defining a variable X_t which can take as a value any of the attractors (S, P1, P2, S1, S2, and C) at each time t. Then, given the six attractors of interest, we would like to estimate the transition matrix P, with the transition probabilities p_{ij} of transition from attractor i to attractor j as components. This matrix can be estimated by iterating Eq. 2 and following the algorithm described in Subheading 2.4.2. Alvarez-Buylla and collaborators [4] followed a similar approach, and estimated the matrix P shown in Table 1. This matrix was estimated using a value of 0.01 for the probability of error ξ in Eq. 2.

We follow the temporal evolution of the probability of reaching each attractor by iterating Eq. 4 using as P the matrix just estimated (*see* Table 1). However, as mentioned in Subheading 2.4.3, it is necessary to specify an initial distribution $p_X(0)$, which defines what fraction of the whole cell population corresponds to each of the cell-types (S, P1, P2, S1, S2, and C) at the initial time of the

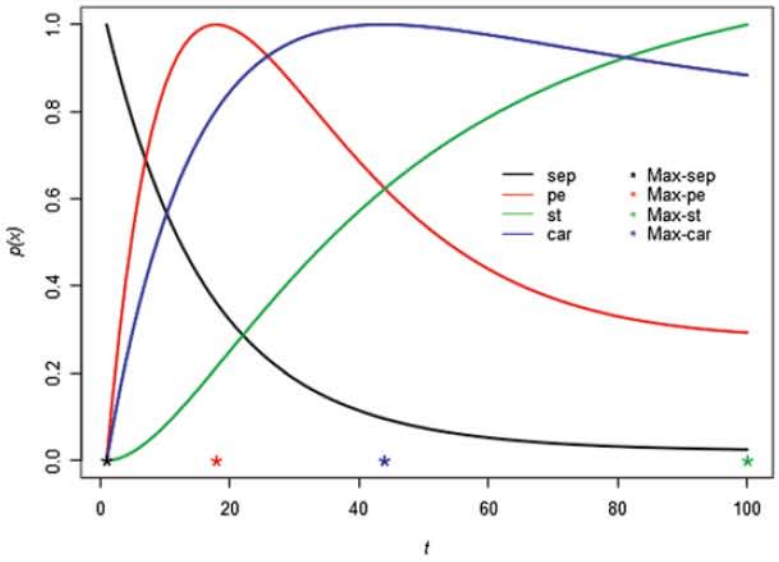


Fig. 8 Temporal sequence of cell-fate attainment pattern under the stochastic Boolean GRN model. Maximum relative probability p of attaining each attractor, as a function of time (in iteration steps). The value of the error probability used was $\xi = 0.01$. *Stars* mark the time when maximal probability of each attractor occurs. The most probable sequence of cell attainment: sepals, petals, carpels, and stamens

simulation. Since sepal primordial cells are the first to attain their fate in flower development, we use as an initial distribution a vector in which the value corresponding to the fraction of sepal cells is set to 1 and all the other values are set to zero; this is $p_X(0) = (1,0,0,0,0,0)$, where the order of the values is (S, P1, P2, S1, S2, and C). Thus, initially, all of the population of cells within a floral primordium is in the sepal attractor. Then, Eq. 4 can be iterated to follow the changes in the probability of reaching each one of the other attractors over time, given that the entire system started in the sepal configuration. The resulting normalized trajectories for the case in point are shown in Fig. 8 (*see* **Note 13**). The graph clearly shows how the trajectory for each of the attractor's probability reaches its maximum at a given time. One star for each of the attractors was drawn in the graph just above the x-axis at the time when its maximal probability occurs. In accordance with biological observations, the results show that the most probable sequence of cell attainment is: sepals, petals, and the stamens and carpels almost concomitantly.

The results presented here were calculated using just one value for the probability of error ($\xi = 0.001$). In the work of Alvarez-Buylla and collaborators [4], it was shown that the system exhibited a sequence of transitions among attractors that mimics the sequence of gene activation configurations observed in real flowers for a level of noise (value of ξ) of around 0.5–10 % (*see* **Note 11**).

The nonintuitive, constructive role of moderated noise perturbing the dynamics of nonlinear systems is a well-known phenomenon in physics [37]. Currently, there is a growing interest in understanding the interplay between noise and the nonlinearity of biological networks [38]. Using the model formalism presented here, Alvarez-Buylla and collaborators concluded that the stereotypical temporal pattern with which floral organs are determined may result from a stochastic dynamic system associated with a highly nonlinear GRN [4]. In the light of these findings, the modeling framework exposed in this section constitutes a simple approach to understanding morphogenesis, providing predictions on the population dynamics of cells with different genetic configurations during development.

2.5 Approximation to a Continuous GRN Model

Boolean GRNs have been useful to study the complex logic of transcriptional regulation involved in cell differentiation because it seems that the qualitative topology of such networks, rather than the detailed form of the kinetic functions of gene interactions, rule the attractors reached. However, for some further mathematical developments and also for studies of the detailed behavior of GRN dynamics, the differences in genetic expression decay rates, threshold expression values, saturation rates, and other quantitative aspects of GRNs can become very relevant. These aspects of GRNs cannot be contemplated by a discrete approach. Hence, it becomes necessary to investigate also continuous representations of GRN dynamics. Several studies reviewed here show that such continuous approximations of the discrete GRNs lead to novel predictions, but at the same time recover consistent results with those arising in the Boolean framework.

2.5.1 Deterministic Approach

Several approaches have been used to describe the Boolean GRN as a continuous system. A well-known scheme is the piecewise linear Glass dynamics of the network [39]. This model is based on a set of differential equations in which each continuous variable x_i, representing the level of expression of a given gene, has an associated discrete variable that represents the state of expression of that gene. This is accomplished by introducing the discrete variables $\hat{x}_i$ defined as $\hat{x}_i = H(x_i - \theta_i)$, where θ_i represents a threshold, and $H(x)$ is the Heaviside step function: $H(x) = 1$ if $x > 1$, and $H(x) = 0$ if $x < 1$. This definition implies that gene n displays a dichotomic expression driven by a more gradual continuous dynamics. The piece-wise continuous Glass dynamics of the GRN is described by

$$\frac{\mathrm{d}x_i(t)}{\mathrm{d}t} = \mu\left[f_i\left(\hat{x}_1(t),\ldots,\hat{x}_k(t)\right) - x_i(t)\right] \tag{5}$$

where f_i are the input functions of the discrete Boolean model, and $\mu = 1/\tau$ is the relaxation rate of the gene expression profile. Within this description, the microscopic configuration of the GRN

at a given time is described by the set of continuous values $\{x_1(t),\ldots,x_k(t)\}$; this set induces in turn the set of corresponding discrete values $\{\breve{x}_1(t),\ldots,\breve{x}_k(t)\}$ as the Boolean configuration of the network. The equilibrium states of the GRN that determine a given phenotype may be obtained from the condition $dx_i/dt=0$, which leads to

$$x_i^S = f_i\left(\breve{x}_1^S(t),\ldots,\breve{x}_k^S(t)\right) \tag{6}$$

independently of the value of the relaxation rate. Even when the Boolean input functions f_i are the same in the discrete and continuous approaches, there are infinitely many microscopic configurations compatible with the same Boolean configuration, and the discrete model of the GRN and the corresponding continuous piece-wise linear model are not necessarily equivalent, since the attractors of the two models can be different. However, numerical simulations to study the GRN for floral organ differentiation in *A. thaliana*, show that the Glass dynamics generate exactly the same ten fixed-point attractors obtained in the Boolean model, although the size of the corresponding attraction basins may display some variation [4].

An alternative approach consists in considering that the input functions display a saturation behavior characterized by a logistic or a Hill function, usually employed in biochemistry to describe ligand saturation as a function of its concentration. In the first case, the input associated to node i may be included in the form

$$\Theta\left[f_i(x_1,,\ldots,,x_k)\right]=\frac{1}{1+\exp\left[-b_i\left[f_i(x_1,,\ldots,,x_k)-\in_i\right]\right]} \tag{7}$$

where $\in_i$ is a threshold level (usually $\in_i=1/2$), and b_i the input saturation rate. It may

be easily seen that for $b_i>>1$, the input function becomes a Heaviside step function:

$$\Theta\left[f_i-\in_i\right]\to H\left[f_i-\in_i\right], \tag{8}$$

and thus displays a dichotomic behavior (in practice this may be achieved for, e.g., $b_i>10$). This approach has been employed, for example, in the modeling of the GRN for differentiation of Th cells of the immune response by Mendoza and Xenarios [40], or in the study of floral organ specification in *A. thaliana* [1].

On the other hand, Hill-type inputs of GRNs have been employed in a number of investigations on biological development and differentiation (*see* the review in ref. 41). They have the following structure:

$$\Xi^{(n)}\left[f_i\right]=\frac{A_i\left(f_i\right)^n}{\left(\in_i\right)^n+\left(f_i\right)^n}, \tag{9}$$

with the parameter n, an integer number, and A_i the maximum asymptotic value attained by the input. The latter approach was used by Zhou et al. [42], to model pancreatic cell fates; and by Wang and coworkers [43] to study myeloid and erythroid cell fates. The approximation to be used depends on the nature of the problem under study. In fact, the GRN inputs could be described also by any set of polynomial functions that reflect the biological interactions of the network.

Another approach that can be used to translate the logical into continuous functions involves the use of "fuzzy logics" proposed by L. A. Zadeh [44] to study systems that do not follow strictly 1 or 0 truth-values. This is achieved by using the following rules

$$\begin{aligned} & x_i(t)\ and\ x_j(t) \to \min\left[x_i(t), x_j(t)\right] \\ & x_i(t)\ or\ x_j(t) \to \max\left[x_i(t), x_j(t)\right] \\ & not\ x_i(t) \to 1 - x_i(t) \end{aligned} \quad . \tag{10}$$

Here, the operators, min and max mean to choose between the minimum and maximum values of the functions x_i and x_j at a given time t. It can be shown that these rules lead to a Boolean algebra [1]. One possible disadvantage of this proposition is that it involves only piece-wise differential functions. Another possibility is to consider the following algorithm:

$$\begin{aligned} & x_i(t)\ and\ x_j(t) \to x_i(t) \cdot x_j(t) \\ & x_i(t)\ or\ x_j(t) \to x_i(t) + x_j(t) - x_i(t) \cdot x_j(t). \\ & not\ x_i(t) \to 1 - x_i(t) \end{aligned} \tag{11}$$

The structure of the expressions associated to the logical connectors "and" and "not" is obvious, while the expression for "or" is derived by substituting such expressions into De Morgan's law: not(x_i or x_j) = (not x_i) and (not x_j). As before, it may be straightforwardly checked that these rules define a Boolean algebra. For example, a logic input like

$$f_1 = \left(x_1\ or\ x_2\right) and\ not\left(x_3\right)$$

would read:

$$f_1 = \left(x_1 + x_2 - x_1 \cdot x_2\right)\left(1 - x_3\right).$$

We now proceed to write the equation for the GRN continuous dynamics. By assuming that the source of gene activation can be characterized, for example, by a logistic-type behavior, we may introduce the following set of differential equations:

$$\frac{\mathrm{d}x_i}{\mathrm{d}t} = \Theta\left[f_i\left(x_1,,\ldots,,x_k\right)\right] - \mu_i x_i \tag{12}$$

where $\mu_i = 1/\tau_i$ represents the expression decay rate of node i of the GRN. Notice that within this approach we consider that, in gen-

eral, each gene may have its own characteristic decay rate. This assumption introduces further richness into the description, as a hierarchy of times of genetic expression may define alternative routes to cell fates. In particular, notice that the steady states of the GRN, given by the condition $dx_i/dt = 0$, lead to the expression

$$x_i^S = \frac{1}{\mu_i}\Theta\left[f_i\left(x_1^S, ,\ldots, , x_k^S\right)\right]. \tag{13}$$

Taking into account that the node inputs are defined by logical sentences with a Boolean architecture, then the attractor set obtained in this case is equivalent by construction to the set derived in the discrete Boolean approach. Thus, if a given attractor arising in the discrete Boolean approach has an expression pattern like $\{1,0,0,1,1,\ldots\}$, the corresponding pattern in the continuous approach would have the structure $\{1/\mu_i, 0, 0, 1/\mu_4, 1/\mu_5, \ldots\}$, so that they become identical when $\mu_i = 1$ (with the possible exception of some isolated attractors). The consideration of the several relaxation rates for gene expression dynamics introduces an important difference with respect to Glass dynamics. For example, in the case that a gene has a large decay rate, corresponding to $\mu_i >> 1$, then $x_i^S \to 0$, and the expression pattern would differ with that arising when $\mu_i = 1$. Then, the dynamic behavior of a gene with a large decay rate (short expression time) would be equivalent to an effective mutation associated to lack of functionality. Similarly, the case $\mu_i << 1$ would correspond to an over-expression of that gene. We conclude that the gene expression dynamics is not only regulated by the GRN interactions topology, but also by the hierarchy of relative expression times of its components.

On the other hand, the system also may acquire very different behaviors depending on the value of the saturation rate. As mentioned before, for $b_i >> 1$, the input function becomes a Heaviside step function. In the case, $b_i = 1$, the input function would show a softer behavior. It turns out that in this latter case the attractor set may change drastically with respect to that obtained in the Boolean-like case. This plasticity could be employed to study regulatory systems with a hybrid functionality consisting of transcriptional regulatory logics that are well described with Boolean GRN, and external or coupled signaling transduction pathways that have continuous behaviors and which can impact the dynamics of some of the GRN components.

3 Notes

1. A developmental module incorporates a set of necessary and sufficient molecular components for a particular cell differentiation or morphogenetic process. It is considered a module because it is largely robust to initial conditions and it attains

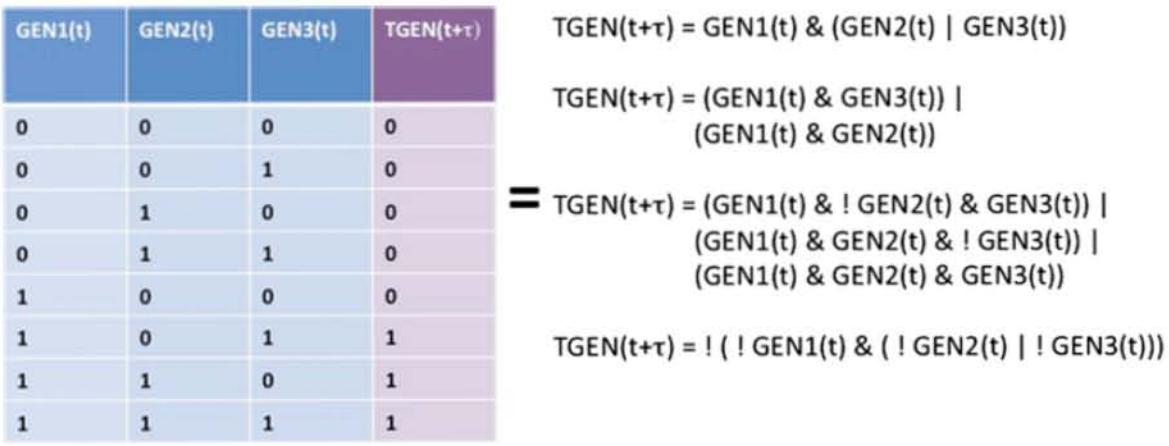


Fig. 9 Equivalence between truth tables and logical statements. As observed each truth table have many equivalent logical statements while each logical statement is represented by a unique truth table

certain attractors robustly. The uncovered GRN underlying the ABC patterns of gene activation and the early subdifferentiation of the flower meristem into four concentric regions or primordial floral organ cells, thus constitutes a developmental model. Other developmental modules involved in flower development could be those involved in: the cellular subdifferentiation of each one of the floral organ primordia during organ maturation, determining floral organ number and spatial disposition, in the dorso-ventrality or shape of floral organs, ovule maturation, etc.

2. In the table that formalizes the experimental data, if the gene or protein is expressed register a "1," and if not a "0." If some components have expression patterns with cyclic behavior, they could be part of cyclic attractors. In some cases, a discrete network with more than two activation states can be postulated if deemed necessary. Quantitative variation in expression levels can be also incorporated later in a continuous model approximated from the discrete one.
3. Several other algorithms exist to numerically find the attractors of a Boolean Network in an efficient way. For examples, *see* ref. 20.
4. It is important to keep in mind that the "AND" and "OR" logical operators can be interconverted. For instance, the logical statement "GEN1 AND GEN2" is equivalent to the logical statement "NOT (NOT GEN1 OR NOT GEN2)." Because of this, most truth tables (except the simplest ones, like the constants) have many equivalent logical statements. Consequently, each Boolean function can be formalized as a unique truth table, but can be described with one or many equivalent logical statements (Fig. 9).
5. Sometimes, the experimental information is not enough to completely define the Boolean functions. For example, in La

Complete characterized Boolean function

GEN1(t)	GEN2(t)	TGEN(t+τ)
0	0	0
0	1	0
1	0	0
1	1	1

Incomplete characterized Boolean function

GEN1(t)	GEN2(t)	TGEN(t+τ)
0	0	*
0	1	0
1	0	0
1	1	1

Fig. 10 Complete and incomplete characterized Boolean functions. While in complete characterized Boolean functions the value of TGEN in all row of the truth tables is specified, in incomplete characterized Boolean functions in one or more rows of the truth table is not specified. Incomplete characterized Boolean function can be the result of missing information data, asynchrony or environmental perturbations and can be resolved with different approaches as explained in the main text

Rota and collaborators [19] Boolean functions were first generated considering only confirmed direct molecular interactions. However, gaps in the experimental information precluded the generation of a unique set of Boolean function determining the GRN. Consequently, they predicted possible interactions by looking for consensus binding sites in the promoters of the included nodes and introducing some speculative hypothesis of molecular interactions.

For example, imagine that TGEN expression disappears when you generate single loss-of-function alleles of GEN1 and GEN2, while TGEN expression is promoted if we over-express both GEN1 and GEN2. Consequently, we conclude that GEN1 and GEN2 are both positive and necessary regulators for TGEN expression. However, this experimental data do not say anything about what happens to TGEN expression in the simultaneous absence of GEN1 and GEN2. In such a case we would have an incompletely characterized Boolean Function (Fig. 10). Such incompletely characterized Boolean functions can also appear due to asynchrony and interactions with the environment [45]. The inclusion of asynchrony in the model provides a more realistic description of our system, while environmental inputs influence is pervasive in biological systems. Hence, the incorporation of incomplete Boolean functions in a model is an instrumental tool. There are many ways to approach this problem: we could test all possible Boolean functions (as in ref. 19), introduce asynchrony in our model, give a probability to each possible Boolean function, or even directly work with incomplete Boolean functions. Several free software programs are capable of considering asynchrony, probabilities for different logical functions or can work with incomplete Boolean functions, such as ANTELOPE [45] and BoolNet [46].

6. Sometimes we cannot represent the available experimental data with a Boolean formalism because we need more values to represent our nodes' activity. For example, imagine that GEN1 differentially affect TGEN in the loss-of function, when normally expressed and when over expressed. This can be resolved replacing the Boolean formalism with a multivalued or a continuous approach. In a multivalued approach, the nodes can take as many values as necessary. In the last example, we could allow GEN1 to have three values, namely, 0 when is OFF, 1 when is normally expressed and 2 when is over expressed. It is important to note that a Boolean formalism can be approximated to a continuous one as was explained in the last section of this paper. For example, Espinosa-Soto and collaborators [7] initially followed a multivalued modeling approach, which was later shown to yield the same qualitative results when transformed into a Boolean system [17]. Similar situations have been documented when transforming a continuous into a Boolean model (e.g., [6, 47]). Currently some software applications allow the analysis of discrete multivalued networks (e.g., GINSIM) [48].
7. As mentioned above, sometimes the experimental information is not enough to generate the Boolean function. We can also find contradictory information linked to particular gene interactions. For example, one author may report that GEN1 positively regulates TGEN, while another one may report that GEN1 is a negative regulator of TGEN. In cases like this, models are extremely helpful, even when they could be considered incomplete. With models we can test both suggestions in a fast and cheap way. The result that better reproduces the experimentally observed system's behavior should be considered the most likely hypothesis. For example, in La Rota and collaborators [19] GRN model of sepal primordium they generated multiple sets of Boolean functions describing their GRN and selected those that recovered the expected attractors and mutant phenotypes. At other times GRN models can be also used to explain apparent contradictions or disputes concerning the interpretation of experimental data.
8. There are several free software packages to recover the attractors and basins of attraction of Boolean GRN, including ANTELOPE [45], GINSIM [48], BoolNet [46], Atalia [12], GNbox [49], GNA [50], and BioCham [51].
9. It is important to note that recovering the expected attractors when the mutants are simulated does not guarantee that the model is correct, because networks with different topologies can sometimes reach the same attractors [52]. However, we can assure that a GRN model that is unable to reproduce all mutants is incorrect.

10. Although stochasticity in Boolean models of GRNs is commonly modeled using the SIN model (*see* Subheading 2.4.1), another method called the stochasticity in functions (SIF) has been introduced recently. The objective of this method is to model stochasticity at the level of biological functions (i.e., Boolean functions in the GRN), and not just by flipping the state of a gene as in the SIN model (for details *see* refs. 20, 32).
11. It could be the case that interesting, nontrivial behaviors may be uncovered just at certain levels of the error probability ξ (e.g., noise). Thus, as customary in numerical explorations, it is necessary to test different values of ξ. However, one expects generic, robust behavior to be observed under a relatively wide range of noise levels. Moreover, the stochastic modeling of GRN can thus be useful to make inferences concerning the range of noise levels that are experienced in particular developmental systems under study.
12. When trying to iterate Eq. 4, make sure that the order in which the position corresponding to each attractor state in the initial vector $p_X(0)$ is the same as the one for the columns in the transition the matrix P. In other words, if the fraction of cells in attractor A is specified in the position i of the initial vector, the row i of the transition matrix should correspond to the probabilities of transition from attractor A to the other attractors.
13. It can be the case that the heights of the trajectories, which correspond to the temporal evolution of the probability of being in each attractor, differ considerably. This is to be expected; given that the basins of the different attractors vary in size, and so do their absolute probabilities. One way to transform the data in order to obtain a graph where the heights of the trajectories are of comparable size is to normalize each probability value with respect to the maximum of each attractor's curve (e.g., dividing the probability value by the maximum value). We followed this approach to obtain the graph in Fig. 8, where also the trajectories corresponding to attractors *se1* and *se2*; and *st1* and *st2* where respectively added to obtain only one trajectory for the attractor *se* and one for *st*. However, it is important to note that, as we are interested in the temporal order in which the attractors reach its maximum probability, this normalization process is not necessary. The order of appearance of the maximum value of the probability of each attractor in the original simulated trajectories would be the same as the one observed in the normalized trajectories. The normalization step just allows us to obtain a clearer graph. In the graph in Fig. 7, we draw one star for each of the attractors just above the x-axis at the time when its maximal probability occurs. The observed pattern is exactly the same in the simulated trajectories before the normalization.

References

1. Villarreal C, Padilla-Longoria P, Alvarez-Buylla ER (2012) General theory of genotype to phenotype mapping: derivation of epigenetic landscapes from N-node complex gene regulatory networks. Phys Rev Lett 109(118102):1–5
2. Alvarez-Buylla ER, Balleza E, Benítez M, Espinosa-Soto C, Padilla-Longoria P (2008) Gene regulatory network models: a dynamic and integrative approach to development. SEB Exp Biol Ser 61:113–139
3. Alvarez-Buylla ER, Azpeitia E, Barrio R, Benítez M, Padilla-Longoria P (2010) From ABC genes to regulatory networks, epigenetic landscapes and flower morphogenesis: making biological sense of theoretical approaches. Semin Cell Dev Biol 21(1):108–117
4. Alvarez-Buylla ER, Chaos A, Aldana M, Benítez M, Cortes-Poza Y, Espinosa-Soto C, Hartasánchez DA, Lotto RB, Malkin D, Escalera Santos GJ, Padilla-Longoria P (2008) Floral morphogenesis: stochastic explorations of a gene network epigenetic landscape. PLoS One 3(11):e3626
5. Mendoza L, Alvarez-Buylla ER (1998) Dynamics of the genetic regulatory network for *Arabidopsis thaliana* flower morphogenesis. J Theor Biol 193(2):307–319
6. Albert R, Othmer HG (2003) The topology of the regulatory interactions predicts the expression pattern of the segment polarity genes in *Drosophila melanogaster*. J Theor Biol 223(1): 1–18
7. Espinosa-Soto C, Padilla-Longoria P, Alvarez-Buylla ER (2004) A gene regulatory network model for cell-fate determination during *Arabidopsis thaliana* flower development that is robust and recovers experimental gene expression profiles. Plant Cell 16:2923–2939
8. Azpeitia E, Benítez M, Vega I, Villarreal C, Alvarez-Buylla ER (2010) Single-cell and coupled GRN models of cell patterning in the *Arabidopsis thaliana* root stem cell niche. BMC Syst Biol 4:134
9. Albert I, Thakar J, Li S, Zhang R, Albert R (2008) Boolean network simulations for life scientists. Source Code Biol Med 3:16
10. Albert R, Wang RS (2009) Discrete dynamic modeling of cellular signaling networks. Methods Enzymol 467:281–306
11. Assmann SM, Albert R (2009) Discrete dynamic modeling with asynchronous update, or how to model complex systems in the absence of quantitative information. Methods Mol Biol 553:207–225
12. Alvarez-Buylla ER, Benítez M, Corvera-Poiré A, Chaos CA, de Folter S, Gamboa de Buen A, Garay-Arroyo A, García-Ponce B, Jaimes- MF, Pérez-Ruiz RV, Piñeyro-Nelson A, Sánchez-Corrales YE (2010) Flower development. Arabidopsis Book 8:e0127
13. Pelaz S, Tapia-López R, Alvarez-Buylla ER, Yanofsky MF (2001) Conversion of leaves into petals in Arabidopsis. Curr Biol 11(3): 182–184
14. Barrio RÁ, Hernández-Machado A, Varea C, Romero-Arias JR, Alvarez-Buylla E (2010) Flower development as an interplay between dynamical physical fields and genetic networks. PLoS One 5(10):e13523
15. Kauffman S (1969) Homeostasis and differentiation in random genetic control networks. Nature 224:177–178
16. Mendoza L, Thieffry D, Alvarez-Buylla ER (1999) Genetic control of flower morphogenesis in *Arabidopsis thaliana*: a logical analysis. Bioinformatics 15(7–8):593–606
17. Chaos Á, Aldana M, Espinosa-Soto C et al (2006) From genes to flower patterns and evolution: dynamic models of gene regulatory networks. J Plant Growth Regul 25(4):278–289
18. Sanchez-Corrales YE, Alvarez-Buylla ER, Mendoza L (2010) The *Arabidopsis thaliana* flower organ specification gene regulatory network determines a robust differentiation process. J Theor Biol 264:971–983
19. La Rota C, Chopard J, Das P, Paindavoine S, Rozier F, Farcot E, Godin C, Traas J, Monéger F (2011) A data-driven integrative model of sepal primordium polarity in Arabidopsis. Plant Cell 23(12):4318–4333
20. Garg A, Mohanram K, De Micheli G, Xenarios I (2012) Implicit methods for qualitative modeling of gene regulatory networks. Methods Mol Biol 786:397–443
21. Alvarez J, Guli CL, Yu XH, Smyth DR (1992) terminal flower: a gene affecting inflorescence development in *Arabidopsis thaliana*. Plant J 2(1):103–116
22. Shannon S, Meeks-Wagner DR (1991) A mutation in the Arabidopsis TFL1 gene affects inflorescence meristem development. Plant Cell 3(9):877–892
23. Parcy F, Bomblies K, Weigel D (2002) Interaction of LEAFY, AGAMOUS and TERMINAL FLOWER1 in maintaining floral meristem identity in Arabidopsis. Development 129(10):2519–2527

24. Conti L, Bradley D (2007) TERMINAL FLOWER1 is a mobile signal controlling Arabidopsis architecture. Plant Cell 19(3): 767–778
25. Chen L, Cheng JC, Castle L, Sung ZR (1997) EMF genes regulate Arabidopsis inflorescence development. Plant Cell 9(11):2011–2024
26. Liljegren SJ, Gustafson-Brown C, Pinyopich A (1999) Interactions among APETALA1, LEAFY, and TERMINAL FLOWER1 specify meristem fate. Plant Cell 11(6):1007–1018
27. Ratcliffe OJ, Bradley DJ, Coen ES (1999) Separation of shoot and floral identity in Arabidopsis. Development 126(6):1109–1120
28. Gustafson-Brown C, Savidge B, Yanofsky MF (1994) Regulation of the Arabidopsis floral homeotic gene APETALA1. Cell 76(1): 131–143
29. Gómez-Mena C, de Folter S, Costa MMR, Angenent GC, Sablowski R (2005) Transcriptional program controlled by the floral homeotic gene agamous during early organogenesis. Development 132(3):429–438
30. Kitano H (2007) Towards a theory of biological robustness. Mol Syst Biol 3:137
31. Whitacre JM (2012) Biological robustness: paradigms, mechanisms, and systems principles. Front Genet 3:67
32. Garg A, Mohanram K, Di Cara A, De Micheli G, Xenarios I (2009) Modeling stochasticity and robustness in gene regulatory networks. Bioinformatics 25:i101–i109
33. Samoilov MS, Price G, Arkin AP (2006) From fluctuations to phenotypes: the physiology of noise. Sci STKE 2006:re17
34. Hoffmann M, Chang HH, Huang S, Ingber DE, Loeffler M, Galle J (2008) Noise-driven stem cell and progenitor population dynamics. PLoS One 3(8):e2922
35. Eldar A, Elowitz MB (2010) Functional roles for noise in genetic circuits. Nature 467(7312):167–173
36. Balázsi G, van Oudenaarden A, Collins JJ (2011) Cellular decision making and biological noise: from microbes to mammals. Cell 144(6):910–925
37. Horsthemke W, Lefever R (1984) Noise-induced transitions: theory and applications in physics, chemistry, and biology. Springer, Berlin
38. Chalancon G, Ravarani CNJ, Balaji S, Martinez-Arias A, Aravind L, Jothi R, Babu MM (2012) Interplay between gene expression noise and regulatory network architecture. Trends Genet 28(5):221–232
39. Glass L (1975) Classification of biological networks by their qualitative dynamics. J Theor Biol 54:85–107
40. Mendoza L, Xenarios I (2006) A method for the generation of standardized qualitative dynamical systems of regulatory networks. Theor Biol Med Model 3:13
41. Ferrell JE Jr (2012) Bistability, bifurcations, and Waddington's epigenetic landscape. Curr Biol 22:R458–R466
42. Zhou JX, Brusch L, Huang S (2011) Predicting pancreas cell fate decisions and reprogramming with a hierarchical multi-attractor model. PLoS One 6(3):e14752
43. Wang J, Zhang K, Xua L, Wang E (2011) Quantifying the Waddington landscape and biological paths for development and differentiation. Proc Natl Acad Sci 108:8257–8262
44. Zadeh LA (1965) Fuzzy sets. Inf Control 8:338–353
45. Arellano G, Argil J, Azpeitia E, Benítez M, Carrillo M, Góngora P, Rosenblueth DA, Alvarez-Buylla ER (2011) "Antelope": a hybrid-logic model checker for branching-time Boolean GRN analysis. BMC Bioinformatics 12:490
46. Müssel C, Hopfensitz M, Kestler HA (2010) BoolNet—an R package for generation, reconstruction and analysis of Boolean networks. Bioinformatics 26(10):1378–1380
47. von Dassow G, Meir E, Munro EM, Odell GM (2000) The segment polarity network is a robust developmental module. Nature 406(6792):188–192. doi:10.1038/35018085
48. Naldi A, Berenguier D, Fauré A, Lopez F, Chaouiya C (2009) Logical modelling of regulatory networks with GINsim 2.3. Biosystems 97(2):134–139
49. Corblin F, Fanchon E, Trilling L (2010) Applications of a formal approach to decipher discrete genetic networks. BMC Bioinformatics 11(1):385
50. de Jong H, Geiselmann J, Hernandez C, Page M (2003) Genetic network analyzer: qualitative simulation of genetic regulatory networks. Bioinformatics 19(3):336–344
51. Calzone L, Fages F, Soliman S (2006) Biocham: an environment for modeling biological systems and formalizing experimental knowledge. Bioinformatics 22(14):1805–1807
52. Azpeitia E, Benítez M, Padilla-Longoria P, Espinosa-Soto C, Alvarez-Buylla ER (2011) Dynamic network-based epistasis analysis: boolean examples. Front Plant Sci 2:92

Chapter 23

Descriptive vs. Mechanistic Network Models in Plant Development in the Post-Genomic Era

J. Davila-Velderrain, J.C. Martinez-Garcia, and E.R. Alvarez-Buylla

Abstract

Network modeling is now a widespread practice in systems biology, as well as in integrative genomics, and it constitutes a rich and diverse scientific research field. A conceptually clear understanding of the reasoning behind the main existing modeling approaches, and their associated technical terminologies, is required to avoid confusions and accelerate the transition towards an undeniable necessary more quantitative, multidisciplinary approach to biology. Herein, we focus on two main network-based modeling approaches that are commonly used depending on the information available and the intended goals: inference-based methods and system dynamics approaches. As far as data-based network inference methods are concerned, they enable the discovery of potential functional influences among molecular components. On the other hand, experimentally grounded network dynamical models have been shown to be perfectly suited for the mechanistic study of developmental processes. How do these two perspectives relate to each other? In this chapter, we describe and compare both approaches and then apply them to a given specific developmental module. Along with the step-by-step practical implementation of each approach, we also focus on discussing their respective goals, utility, assumptions, and associated limitations. We use the gene regulatory network (GRN) involved in *Arabidopsis thaliana* Root Stem Cell Niche patterning as our illustrative example. We show that descriptive models based on functional genomics data can provide important background information consistent with experimentally supported functional relationships integrated in mechanistic GRN models. The rationale of analysis and modeling can be applied to any other well-characterized functional developmental module in multicellular organisms, like plants and animals.

Key words Gene regulatory networks, Root stem cell niche, Cell differentiation, Attractor, Morphogenesis, System dynamics, Mathematical model, Computational simulation, Network inference, Descriptive model, Mechanistic model

1 Introduction

Mathematical modeling and computational modeling are becoming an indispensable scientific research practice in modern post-genomic biology. The term *systems biology* has been coined to define this new field of study, highly characterized by its fuzzy disciplinary boundaries. The *systems* perspective to biology embraces the notion of biological behavior as resulting from the collective action of

Jose M. Alonso and Anna N. Stepanova (eds.), *Plant Functional Genomics: Methods and Protocols*, Methods in Molecular Biology, vol. 1284, DOI 10.1007/978-1-4939-2444-8_23,

multiple interacting components at different temporal and spatial scales and levels of organization. Collective behavior emerges from the component interactions themselves and not only from the specific function of the individual components of a given complex system. Multicellular development includes several such collective processes involving molecular genetic components that lead to cell growth, proliferation, and differentiation, and to the eventual emergence of spatial and temporal structural morphogenetic patterns. All these dynamical processes are, to a great extent, self-organized and thus occur as unorchestrated choreographies that can be understood in terms of specific properties of networks or dynamical patterning modules of different nature [1–4]. The study of collective phenomena in biological systems, however, requires approaches that go beyond the discovery and description of individual molecular components [5, 6]. Uncovering how dynamical behavior emerges and is robustly maintained, from the genetic and non-genetic components and their interactions, requires the use of mathematical/computational models [6–9]. In this chapter, we show how these formal tools enable the integration of molecular genetic data into network-based models.

The ongoing genomic revolution has been quite successful in uncovering a fairly complete set of molecular components at different levels of regulation and for multiple organisms [10–13]. At the same time, developmental genetic studies have successfully characterized sets of molecular regulators known to be tightly associated with specific developmental processes, and with the establishment of morphogenetic patterns [14–16]. In post-genomic biology there is an increasing need to transcend the reductionist modes of explanation, to go beyond the traditional enumeration and the bookkeeping description of molecular processes and components, and to integrate this knowledge into explanatory models [5, 17, 18]. Towards this goal, we can distinguish two important questions: (1) given a set of known molecular players, how can we gain insights into their regulatory interactions; and (2) once a set of molecules and their interactions are known, how can we study the associated dynamic behavior and, ultimately, the phenotypic manifestation of such a molecular regulatory system. In this chapter, we show how to approach these questions within the context of the practical implementation of gene regulatory network (GRN) models.

GRN models are considered as one of the most powerful tools for the study of complex molecular systems [2, 4, 7]. A GRN is composed of a given set of molecular players (e.g., genes, proteins) and a given set of interactions among them, which represent regulatory influences. Then, for the case of GRNs, the question (1) above refers, more precisely, to the process of inferring these interactions from some source of experimental data [19–21]. Question (2) above implies a mechanistic perspective: the use of additional information and assumptions about underlying processes driving

the dynamical behavior in order to simulate it and to uncover the consequences of the dynamical interplay [6]. From the modeling point of view, these two tasks are associated with two different approaches. (1) Data-based *descriptive* models are used to postulate putative regulatory interactions among molecular players through the quantitative descriptions of the observed relationships among a set of measured variables. On the other hand, (2) *mechanistic* dynamical models are used to represent, in a quite useful simplified manner, specific processes underlying cell behavior, using for this well-posed descriptive equations or computer-based encoded systems knowledge [22]. In the latter case, the resultant models enable the study of how cell behavior changes over time, as well as the long-term consequences of the underlying dynamical processes.

The descriptive (*statistical*) approach is commonly used as a way to make sense of large-scale genomic data [23, 24]. On the other hand, the mechanistic perspective is widely applied to small or moderate-order well-characterized biological processes [2]. Given that genome-scale networks are composed of multiple structural and functional modules [25–28], others and we have proposed to use GRN models to discover robust modules and explore their dynamic behavior [29–31]. Following this line of research, in this chapter we contrast the descriptive and mechanistic approaches taking as an illustrative example a recently well-characterized GRN model: the GRN involved in *Arabidopsis thaliana* Root Stem Cell Niche (SCN) developmental dynamics [32, 33]. Using this developmental module as an example, we show: (1) how a data-based, descriptive approach can be applied to propose putative gene interactions that later can be included in mechanistic GRN models; (2) how a dynamical GRN model is constructed from published molecular experimental data; (3) the common steps followed in the dynamic analysis of a GRN mechanistic model; and (4) a comparison between the inferred descriptive GRN model and the well-characterized mechanistic dynamic GRN model.

1.1 Definitions

Network modeling in post-genomic biology is a diverse practice. Different, well-established traditions exist within the mathematical and physical sciences, where terms and definitions are commonly adopted dependent on the context. In multidisciplinary fields such as systems biology and integrative genomics, however, such distinctions get blurred. The problem is particularly acute in molecular network modeling: computer scientists, statisticians, engineers, physicists, and mathematicians are all trying to approach the problem making important contributions [24, 34–36]. It is difficult to devise a consensus within such diversity. Aware of this problem, we start by conceptually distinguishing between the two general modeling traditions, namely a *descriptive* vs. a *mechanistic* modeling approach. For each case, we define key terminology to be used in

the sections that follow. Although we focus the discussion on GRN models, the comparison in this section concerns the general practice of mathematical/computational modeling. In particular, for each modeling perspective, we define general modeling concepts such as *validation*, *prediction*, and *explanation* (*see* Table 1).

1.1.1 Descriptive Models

A descriptive model is a quantitative summary of the observed relationships among a set of measured variables [22]. In the case of GRN modeling, the variables commonly correspond to genes whose activity is measured by quantifying gene expression. Functional genomic data (e.g., microarray or next-generation sequencing (NGS) data) are commonly used as the set of measurements [19].

Goals: The main goal of descriptive, inferential approaches is to discover new *knowledge*. In general, descriptive models aim at finding

Table 1
General modeling concepts

Descriptive modeling	
Model	A mathematical expression or computer algorithm that relates the values of one or more *responsive* (dependent) variables with the values of a set of *predictor* (independent) variables.
Prediction	Calculated values of the responsive variables by taking specific values of the predictor variables as input to the model.
Explanation	A predictor variable x is said to *explain* a responsive variable y if the predicted values for y are in agreement (to a certain degree) with the observed values in a particular dataset comprising empirical values of x and y.
Validation	The practice of testing the performance of a model by testing its predictive power using an independent dataset.
Causal attribution	It is not possible to postulate the reasons why a certain quantitative relationship embedded in the model is able to *explain* one variable in terms of the other—"*correlation does not imply causation*."
Mechanistic modeling	
Model	Set of equations or computer code that describe how simplified properties of a real-world entity (system) change over time as a result of specific underlying processes.
Prediction	Forecasting the future properties of the system or their long-term behavior.
Explanation	The processes considered in the model account for the observed system behavior.
Validation	The practice of contrasting model predictions with experimental observations of the real-world entity.
Causal attribution	The predicted behavior results from the underlying *causal* processes considered in the model. The model is built by explicitly considering the processes that produce our observations.

novel hypotheses regarding the functional influences among molecular components amidst the mass of high-throughput data [23]. In the case of GRN models, this corresponds precisely to finding putative function-specific network nodes and edges.

Main assumptions: The reasoning is based on the idea that molecular components that share discernible patterns in high-throughput data sets also share experimentally testable biological (functional) relationships.

Main limitations: (1) A descriptive model says nothing about *why* the variables are related the way they are (*see* **Note 1**). (2) We can only be confident that the relationships apply to the conditions (e.g., samples) where the data come from (*see* **Note 2**). It might apply to other conditions, for example to the same tissue, or even to other tissues, but it might not.

Conclusions to draw: The connected nodes (genes) show certain coordinated statistical activity through the sample conditions included in the data set (*see* **Note 3**). Subsets of molecules participating in similar biological processes, even if they do not have physical interactions, can be uncovered with these models. However, the observation of correlated behavior does not necessarily imply a functional relationship (causalities are not always easy to discern). The results should be taken as one source of inconclusive evidence—useful to be integrated with further analysis, nonetheless. We must point out that diverse applications that follow a descriptive approach have been integrated recently in the analysis of plant transcriptomes [37, 38].

1.1.2 Mechanistic Model

A mechanistic, dynamic model is a simplified representation of some real-world entity, in terms of descriptive equations or computer-based encoded systems knowledge [22]. The model is called *dynamic* because it describes how system properties change over time. A dynamic model is *mechanistic* because it is built by explicitly considering the processes that produce our observations (i.e., the involved processes are considered in term of the workings of coupled individual components). Relationships between variables emerge from the model as the result of the underlying processes. In the case of GRNs, the process of interest is developmental dynamics, i.e., the establishment of the patterns of cellular differentiation and structural morphogenesis [4, 7].

Goals: The main goal of the mechanistic approaches is scientific understanding [17, 22, 39]. More specifically, answering question such as: How do we create understanding out of validated bits of knowledge? Can processes A and B account for pattern C? Which of several contending sets of assumptions is best able to account for the data? Given that processes A and B occur, what consequences do we expect to observe? Where are the holes in our understanding? (*see* **Note 4**).

Main assumptions: In a mechanistic model, the postulated underlying processes, thought to be driving the system's observed behavior, effectively constitute assumptions. These assumptions should reflect the current state of domain knowledge. In the case of GRNs, it is generally postulated that the time-dependent behavior of the activity of each gene is driven by the coordinated behavior of the genes regulating it, which are in turn also subject to regulation. The overall result of such complex network of mutual regulatory interactions is a restrictive behavior: the present activity state of all the genes in the network, and the regulatory interactions among them, determine the future activity state.

Main limitations: (1) Identifying which state variables and processes are important for your modeling purposes is not trivial. Thus, the construction of mechanistic models is a time-consuming process. (2) The available knowledge upon which the model is constructed is essentially incomplete. In the case of GRNs, it is frequently the case that certain molecular players and key regulatory interactions have not been characterized by the time the model is constructed. But the GRN construction process and modeling is useful to identify and evaluate such gaps in experimental knowledge (*see* **Note 5**); this is one of the most important advantages of the system dynamics approach.

Conclusions to draw: The observed behavior is a direct consequence of the underlying processes considered in the model. The observed behavior resulting from simulated interventions can constitute *predictions* (*see* Table 1). For example, in GRN dynamical models the expression profile represents or correlates with particular cellular phenotypes (*see* Table 2). The modeled regulatory interactions restrict the permissible behavior of the time-changing expression profile, and also determine the existence of certain stable, time-invariant expression profiles. Multiple studies have shown that these stable configurations correspond to those characterized in several cell types but for which a mechanistic and dynamical explanation was lacking [2, 4, 30]. Therefore, stable cellular phenotypes, as described by gene expression profiles, result from the restrictions imposed by a given GRN. Furthermore, loss- or gain-of-function mutations can be easily simulated as controlled interventions in the model. The effect of these simulated interventions on the observed stable expression profiles can be useful to validate the model derived from the considered wild-type (wt) constraints, or can also constitute predictions subjected to experimental *validation* (*see* Table 1).

1.1.3 Descriptive vs. Mechanistic

A dynamic model is built up from descriptive equations representing the processes thought to account for the patterns observed in the given data, whereas a descriptive model only represents the patterns themselves. Do these two strategies have to be mutually

Table 2
GRN dynamical model concepts

Concept	Definition
Node	Representation of a molecular species (gene, protein, etc.).
Edge	Representation of a given regulatory interaction.
Node state (variable)	Expression value that a node takes at a certain time.
Network state	Ordered set of node expression values at a certain time.
State space	Set comprising all possible network states.
Attractor	Stable and stationary (time-invariant) network states.
Transitory state	Network states that are not (do not form part) of an attractor (attractor's basin).
Basin of attraction	Set comprising all the initial network states that eventually lead to a particular attractor.
Biologically observable attractor	Gene expression profiles (gene configurations) that have been obtained from experimental assays and reported in the scientific literature for particular cell types.

exclusive? We consider that the integration of descriptive and mechanistic models is a promising, yet rarely applied, approach in post-genomic biology. Incomplete knowledge is a common limitation for the postulation of mechanistic GRN models. On the other hand, the main goal of descriptive models is to uncover new knowledge from high-throughput data, which is quite vast and increasing in post-genomic biology. In our opinion, this distinction can be exploited in order to circumvent the limitations of each individual approach. The *predictions* (*see* Table 1 for definitions) made by following a descriptive approach can be used as a source of knowledge to be integrated into a mechanistic model. In order for this suggested model integration strategy to be useful, however, the descriptive predictions should be accurate. How do we test if this is the case? We approach this issue in Subheading 3.3 below.

In the following sections, we show how to apply both a descriptive and a mechanistic modeling approach taking a well-defined regulatory module as a simple illustrative example.

2 Materials

2.1 Descriptive Approach to GRN Modeling

2.1.1 Data

Arabidopsis thaliana Root Genome-wide GRN: In a recent study, Montes and collaborators applied network inference to publicly available *Arabidopsis thaliana* root microarray samples [45]. They compiled a dataset of microarray samples from the EBI ArrayExpress database based on the following criteria: (1) include only experiments

using the Affymetrix GeneChip ATH1-121501, (2) include only data corresponding to root tissues, (3) exclude samples from ecotypes other than Columbia-0, and (4) exclude transgenic samples (mutant and overexpression lines, and promoter constructs). The final dataset consists of 656 microarray samples. The raw microarray data was pre-processed using the R package *gcrm* to obtain the expression matrix (for details, *see* ref. 45). For illustration purposes, here we use this dataset for the inference exercises. All the inferences shown below are based on the data extracted from this microarray expression matrix.

Arabidopsis thaliana Root Stem Cell Niche (SCN) GRN: In an attempt to explain the robust patterning of the root SCN of *Arabidopsis thaliana* in terms of the dynamics of known molecular regulators, Azpeitia and collaborators recently postulated several GRN dynamical models [32]. The models are grounded on experimental evidence of the interactions among the main molecular regulators of root SCN patterning. We take this prior experimental information as the basis for the models developed in this chapter. In order to have a direct comparison between the inferred (descriptive) and the dynamical (mechanistic) GRN models, we extract from the dataset of Montes and collaborators [45] only the expression data corresponding to the set of molecular regulators considered by Azpeitia and collaborators [32]. In Table 3, we show a summary of the supporting experimental evidence. We consider these characterized interactions as the "real" interactions set, against which all the inferences would be tested. Accordingly, from the complete expression matrix (*see* Subheading 2.1.1) we extracted only the rows corresponding to the set of genes involved in the "real" interactions set. All the inferences are based on this smaller expression matrix.

2.1.2 Software

Correlation calculations: R statistical programming environment (www.R-project.org).

Mutual information based inference: *minet*, R package [47].

Network visualization: R package *Rgraphviz* [48].

2.2 Mechanistic Approach to GRN Modeling

We take the experimental data in Table 3 as the basis to define the list of state variables (genes) and the corresponding set of Boolean rules.

2.2.1 Data

Experimental expression profiles (expected attractors) are extracted from ref. 32.

Mutant phenotypes are extracted from ref. 32.

2.2.2 Software

BoolNet, R package [60].

2.3 Inference Performance

PPC-based co-expression network (Fig. 1).

MI-based co-expression networks (Fig. 2).

"Real" network (Fig. 3).

2.3.1 Data

2.3.2 Software

minet, R package [47].

Table 3
Experimentally supported (real) interactions set

Interactions	Experimental evidence
SHR→SCR	The expression of *SCR* is reduced in *shr* mutants. ChIP-QRTPCR experiments show that SHR directly binds in vivo to the regulatory sequences of *SCR* and positively regulates its transcription.
SCR→SCR	In the *scr* mutant background, promoter activity of *SCR* is absent in the QC and CEI. A ChIP-PCR assay confirmed that SCR directly binds to its own promoter and directs its own expression.
JKD→SCR	*SCR* mRNA expression as probed with a reporter lines is lost in the QC and CEI cells in *jkd* mutants from the early heart stage onward.
MGP–|SCR	The double mutant *jkd mgp* rescues the expression of *SCR* in the QC and CEI, which is lost in the *jkd* single mutant.
SHR→MGP	The expression of *MGP* is severely reduced in the *shr* background. Experimental data using various approaches have suggested that *MGP* is a direct target of SHR. This result was later confirmed by ChIP-PCR.
SCR→MGP	SCR directly binds to the *MGP* promoter, and *MGP* expression is reduced in the *scr* mutant background.
SHR→JKD	The post-embryonic expression of *JKD* is reduced in *shr* mutant roots.
SCR→JKD	The post-embryonic expression of *JKD* is reduced in *scr* mutant roots.
SCR→WOX5	*WOX5* is not expressed in *scr* mutants.
SHR→WOX5	*WOX5* expression is reduced in *shr* mutants.
ARF(MP)→WOX5	*WOX5* expression is rarely detected in *mp* or *bdl* mutants.
ARF→PLT	*PLT1* mRNA region of expression is reduced in multiple mutants of *PIN* genes, and it is overexpressed under ectopic auxin addition. *PLT1* and *2* mRNAs are absent in the majority of *mp* embryos and even more so in *mp nph4* double mutant embryos.
Aux/IAA–|ARF	Overexpression of *Aux/IAA* genes represses the expression of *DR5* both in the presence and absence of auxin. Domains III and IV of Aux/IAA proteins interact with domains III and IV of ARF stabilizing the dimerization that represses ARF transcriptional activity.
Auxin–|Aux/IAA	Auxin application destabilizes Aux/IAA proteins. Aux/IAA proteins are targets of ubiquitin-mediated auxin-dependent degradation.
CLE40–|WOX5	Wild-type root treated with CLE40p show a reduction of *WOX5* expression, whereas in *cle40* loss-of-function plants *WOX5* is overexpressed.

3 Methods

3.1 Descriptive Approach to GRN Modeling

The practice of inference within systems biology is commonly associated with terms such as reverse engineering [37, 21], data-driven modeling [40], or network learning [41]. Here we refer to all these practices as *descriptive* modeling, as they rely on finding

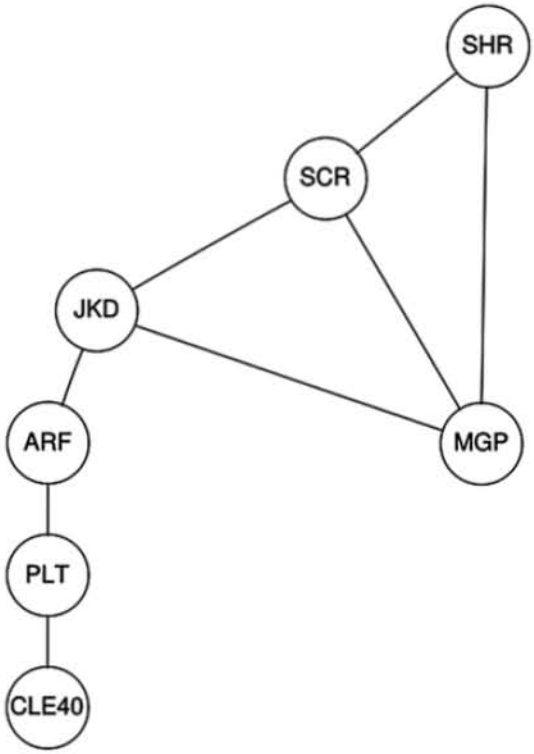


Fig. 1 PPC-based inferred GRN. The *graph* shows the inferred gene interactions among the molecular players included in Table 3. Only those interactions involving a PPC value equal or greater than 0.3 were included in the network (*see* **Note 6**). The inferred GRN qualitatively resembles the real, experimentally supported GRN (*see* Fig. 4)

statistical patterns in the genomic data either at the DNA, mRNA, protein, or metabolic level. Importantly, we do not include here the problem of inferring parameters of mechanistic models from data [42], a practice that may be difficult to classify under the scheme we chose. Multiple statistical models are currently used for network inference purposes [20]. Here we focus exclusively on those models that have been most widely used in plant genomics and systems biology, namely, co-expression networks based on either (1) pair-wise correlation [43, 44], or (2) mutual information criteria [45, 46]. Inference of GRNs by estimating statistical patterns of co-expression is a widely used practice [2, 20].

3.1.1 Pairwise Correlation Co-expression Network

Comparing expression patterns between genes is the basis for constructing a co-expression network [49]. A straightforward definition of a gene co-expression network is a network in which an edge between a given node, say A, and a related node, say B, is added if some measure of similarity between the expression profiles of gene A and gene B exceeds some threshold value, although more stringent algorithms exist (*see* below). One of the most simple and widely used measures of similarity for network construction is the

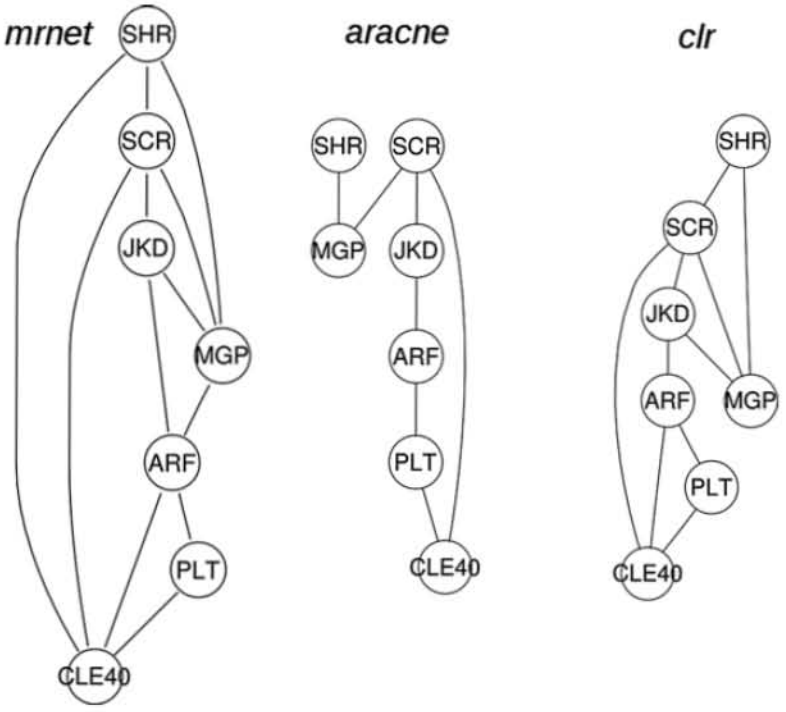


Fig. 2 MI-based inferred GRNs. Graphs of the MI-based inferred GRNs corresponding to each of the algorithms were implemented in the package *minet*. The inferred GRNs are in general more connected than the one based on PPC inference. *CLE40* is a molecular player that was hypothesized to be interacting with *WOX5* (not included because of lack of expression data). *WOX5* in turn interacts with *SCR, SHR*, and *ARF* (*see* Azpeitia et al. [32]). Interestingly, the *mrnet a*lgorithm, which has been shown to perform better than other MI-based algorithms, uncovered co-expression interactions between *CLE40* and the interacting partners of *WOX5*

Pearson correlation coefficient (PCC) [50]. This quite useful approach has been applied several times in plant genomic studies using different expression datasets, and mostly for the analysis of genome-scale networks (*see*, for example refs. 43, 51, 52).

A generic protocol to construct a PPC-based co-expression network for the genes involved in the experimental data summarized in Table 3 would be as follows:

1. A matrix with numbers representing gene expression values is required. In this matrix rows correspond to the genes of interest to be integrated in the network. Columns correspond to the samples where gene expression was measured. We refer to such a matrix as the *expression matrix*. Here, for illustration, we use a data matrix extracted from [45] which corresponds to expression data of the genes summarized in Table 3 (we excluded *WOX5*, as it does not have a unique Affymetrix microarray identifier).

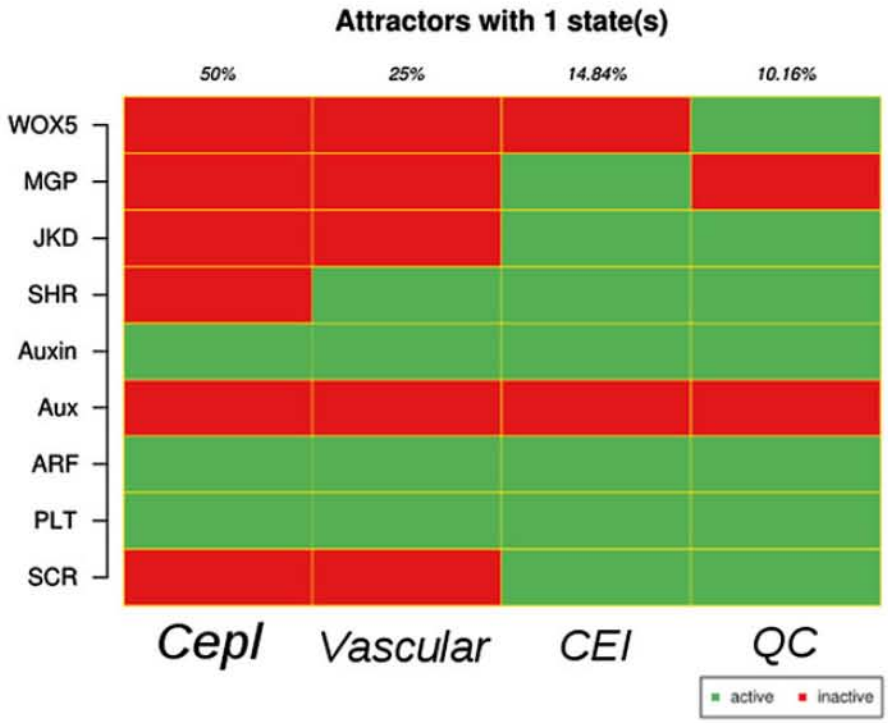


Fig. 3 Obtained attractors of the root SCN GRN. The GRN recovered four fixed-point attractors corresponding to the Root SCN patterning cell types: quiescent center (QC), vascular initials, Cortex–Endodermis initials (CEI), and columella–epidermis–lateral root cap initials (Cepl). In the graph, *green color* indicates expression or gene activation (1), while *red color* indicates no expression or inactivation (0)

2. Given the expression matrix, Pearson correlation coefficient (PCC) values are calculated between pairs of rows (i.e., expression profiles). The function *cor* implemented in the R statistical programming environment can be used for this purpose. Specifically, the expression matrix is given as input to the *cor* function and it automatically calculates PCC values between all possible pairs of rows retrieving a *correlation matrix*, i.e., a matrix whose element *i,j* represents the PCC value between genes *i* and *j*.
3. Given the correlation matrix, an edge is defined between the genes *i* and *j* if the PCC value between them is greater or equal to user-specified threshold value. The complete co-expression network results from defining all gene pairs fulfilling the requirement (*see* **Note 6**).
4. The co-expression network can be plotted using the R package *Rgraphviz* using as input a list of the edges defined to be included in the network.

3.1.2 Mutual Information Network Inference

A very popular inferential approach is based on applying well-established tools from standard information theory [2, 21, 53, 54]. Interactions in these types of inferred co-expression

networks represent a high-degree of statistical dependence between gene expression profiles. These dependencies are typically measured by mutual information (MI) [47]. The adoption of mutual information in network inference is said to circumvent some of the limitations of PPC-based approaches (*see* **Note 7**). Recent studies have shown the utility of MI-based co-expression network inferences for uncovering biological knowledge from plant transcriptomes [45, 46]. Several tools are available for direct implementation of MI-based inferences [47, 55, 56].

Given gene expression data in the form of a gene expression matrix (*see* Subheading 2.1.1), the inference of a MI-based co-expression network consists of two main steps, (1) MI computation and (2) network inference. Thus, a generic protocol infers interactions among Root SCN regulators using the R package *minet* [47], as follows:

1. MI computation: pairwise MI calculations are performed in order to obtain a mutual information matrix (MIM). The function *build.mim* from the *minet* package can be used for this purpose.
2. Network inference: based on the calculated MIM, one of several algorithms is used to select which interactions are included (excluded) to produce a final network. The simplest approach is to choose a threshold MI value, as it was done with the PPC-based network above. However, the *minet* package implements three different algorithms that go beyond the threshold approach in an attempt to reduce the likelihood of inferring indirect interactions, i.e., situations where, for example, a MI value between A and B is high because a third gene C is regulating both A and B (*see* ref. 54 for details). The three algorithms are CLR, ARACNE, and MRNET, and these can be implemented by the respective functions *clr*, *aracne*, and *mrnet* using the previously calculated MIM as input.
3. **Steps 1** and **2** can be applied sequentially using the main function *minet()*. This function implements sequentially all the steps required for the inference, starting directly from the expression matrix and taking the user-selected algorithms as arguments.

We applied the protocols described above to obtain one PPC-based (Fig. 1) and three MI-based co-expression networks (Fig. 2). Importantly, in co-expression networks auto-regulatory interactions are not considered, nor is the directionality of each interaction.

3.2 Mechanistic Approach to GRN Modeling

Dynamic models are diverse, among other things, in terms of the mathematical setting of the model (continuous or discrete time and model variables, deterministic or stochastic, etc.). For simplicity, here we focus on discrete time and discrete state, deterministic

dynamic models. The most widely used GRN model of this type is the Boolean network model [29, 57, 58]. The extension of that dynamic model into more complex models, as well as a more detailed exposition of their analyses, has been reviewed recently by the authors (*see* refs. 7, 59). A dynamic GRN Boolean model has two essential components:

1. A short list of state variables (genes) that are taken to be sufficient for summarizing the properties of interest in the developmental system, and predicting how those properties will change over time. In a Boolean GRN the variables can only attain one of two possible values: *1* if the node is *ON*, and *0* if the node is *OFF*. A *0* node value represents that a gene is not being expressed, while a *1* node value represents that a gene is expressed. These are combined into a *state vector* (in simple terms: a vector is an ordered list of numbers) (*see* Table 2 for definitions).
2. The dynamic equations: a set of equations (or rules) specifying how the state variables change over time, as a function of the current and past values of the state variables (we say that the concerned system is causal and not memory less). In a Boolean model these rules are specified in terms of logical propositions or truth tables (see below).

Thus, a generic protocol to postulate a GRN model for a particular developmental module would be as follows:

1. Define the list of state variables (genes): based on available experimental data, select the set of potential nodes or molecular components that will be incorporated in the GRN model.
2. Define the dynamic equations: collect statements on well-established gene dependencies from literature and express them as Boolean rules or truth tables.
3. Define the "expected attractors": integrate in a Boolean vector the observed expression profiles of the cell-types of interest corresponding to the developmental system being modeled. For this, experimental data concerning the spatiotemporal expression patterns of the genes to be incorporated in the model can be used.
4. Perform a dynamic analysis of the defined GRN model defined in **steps 1** and **2** using a computer-based simulation tool. Identify the stable gene configurations ("simulated attractors").
5. Compare the simulated attractors to the ones observed experimentally (expected attractors; *see* **step 3** above) (*see* **Note 8**).
6. Validate the model by addressing if it recovers the wild-type and mutant (loss- and gain-of-function) gene activation configurations that characterize the cells being considered.

Table 4
Boolean GRN model

List of state variables
X = [SCR, PLT, ARF, Aux, Auxin, SHR, JKD, MGP, WOX5]
Boolean functions
SCR = SHR & SCR
PLT = ARF
ARF = !Aux
Aux = !Auxin
Auxin = !Auxin|Auxin
SHR = SHR
JKD = SHR & SCR
MGP = SHR & SCR & !WOX5
WOX5 = ARF & SHR & SCR & !(MGP & !WOX5)

In the following section, we show a practical implementation of this general protocol using the Arabidopsis root SCN GRN as a simple illustrative example.

3.2.1 Mechanistic Modeling of Arabidopsis Root SCN GRN

Define the list of state variables (genes): Through an exhaustive review of literature, Azpeitia and collaborators identified the set of molecules included in Table 3 as potential members of a developmental module [32]. This set is taken as the list of state variables for the GRN Boolean model (*see* Table 4).

Define Boolean rules: A major advantage of Boolean networks is the fact that natural-language statements can easily be transferred into Boolean representation. The discrete-time Boolean formalism is useful to postulate the set of components and interactions that are necessary and sufficient to recover a particular observed multivariable state (for example, a gene expression configuration). The same logic can be used as well to integrate both molecular genetic and non-genetic components, for example: the effect of mechanical forces, geometric constraints, or chemical components [8, 9]. Here we illustrate this process taking as an example the experimental evidence regarding the functional relationships between the genes SCR and SHR (*see* Table 3).
Natural-language statement 1:

"*The expression of SCR is reduced in shr mutants. ChIP-QRTPCR experiments show that SHR directly binds in vivo to the regulatory sequences of SCR and positively regulates its transcription.*"

Transforming this into a Boolean rule is rather simple:

SCR value after transition depends on SHR, and its value is reduced if SHR is reduced.

Thus, the corresponding transition rule is

$$SCR = SHR$$

Natural-language statement 2:

"*In the scr mutant background promoter activity of SCR is absent in the Root SCN patterning cell types quiescent center (QC) and Cortex-Endodermis initials (CEI). A ChIP-PCR assay confirmed that SCR directly binds to its own promoter and directs its own expression.*"

SCR value after transition depends also on itself, and its promoter activity is reduced if SCR is reduced.

Thus, the transition rule is

$$SCR = SCR$$

In both cases, the regulatory influence is positive. Taken both rules together we obtain the rule:

$$SCR = SHR \,\&\, SCR$$

where & represents the AND operator. The rule means that SCR will be expressed in the future time step if both SHR and SCR are expressed in the present time step.

Following this intuitive transformation process from natural-language statements into Boolean rules or truth tables, one rule for each gene can be postulated. The set of genes with their corresponding Boolean rules completely specifies the Boolean GRN (*see* Table 4).

Define the "expected attractors": Azpeitia and collaborators defined four cell-type expression profiles based on spatiotemporal experimental data from literature sources (*see* Table 5). These profiles are taken as the set of "expected attractors", which the model is expected to recover dynamically as a result of the restrictions imposed by the regulatory interactions encoded in the Boolean rules. Hence such modeling approach enables a mechanistic and dynamical explanation for the observed gene expression configurations.

Analyze GRN model dynamics: Once the set of Boolean rules is specified, these can be loaded directly into the *BoolNet* R package (*see* **Note 9**). This software is able to read in networks consisting of such rule sets, as specified in Table 5, in a standardized text file format (*see* ref. 60). Attractors are stable cycles of states in a Boolean network. As they comprise the states in which the network resides most of the time, attractors in models of GRNs developmental modules are expected to correspond to cellular phenotypes (cell-

Table 5
Gene expression profiles (expected attractors)

Cell type	PLT	Auxin	ARF	Aux/IAA	SHR	SCR	JKD	MGP	WOX5
QC	1	1	1	0	1	1	1	0	1
Vascular initials	1	1	1	0	1	0	0	0	0
CEI	1	1	1	0	1	1	1	1	0
Cepl	1	1	1	0	0	0	0	0	0

type specific expression profiles). The *BoolNet* package is able to identify attractors through the function *getAttractors()*. This function incorporates several methods for the identification of attractors, using as default an exhaustive synchronous search strategy. The identified attractors can then be plotted using the function *plotAttractors()*. We applied these functions to the Root SCN GRN and identified four attractors (*see* Fig. 3).

Comparison of simulated and observed/expected attractors: As expected, the simulated attractors uncovered by the GRN model dynamics (*see* Fig. 3) correspond with the "expected attractors" defined by experimental data (*see* Table 5). This suggests that cell-type specification patterns in the root SCN result from the restrictions imposed by the uncovered GRN developmental module. Defining the expected set of attractors is an indispensable step when building the GRN model, because they are used to validate the GRN. However, it should be clear that the postulation of the Boolean functions is an independent task and, hence, it does not imply circularity.

Simulations of mutant gene knockout and overexpression configurations: For validation purposes, it is straightforward to implement knockout and overexpression simulation experiments within the *BoolNet* package. Specifically, genes can be set to a fixed value (0 for knockout, and 1 for overexpression), and in any calculation on the network this fixed value is taken instead of the value of the corresponding transition function. The function *fixGenes()* takes as input the network, the name of the gene to be perturbed, and the value to be fixed (0 or 1). Then all the other analysis, such as attractors' identification, can be performed over this new perturbed network. Azpeitia and collaborators followed this approach and showed that most predicted alterations to the stable configurations caused by mutant simulations where consistent with known empirical observations [32]. This validates the uncovered dynamical module or set of restrictions as necessary and sufficient to explain the observed gene expression configurations.

3.3 Inference Performance

In the previous sections, we first applied a descriptive approach to GRN modeling in order to infer GRN interactions from gene expression data. As a result, we constructed four inferred GRNs (Figs. 1 and 2). We then described the assemblage and analysis of an experimentally grounded GRN mechanistic model. In this section, we show how to assess the different network inference algorithms. We are interested in knowing if the inferred interactions are consistent with the ones defined based on published molecular functional experimental data. Once a "true" network is defined, there exist well-established tools to assess the performance of the inference algorithms. In this section, we take as a "true" network the one based on well-curated functional molecular genetic data and call it the mechanistic SCN GRN model that integrates the interactions summarized in Table 3. The model is shown in Fig. 4. In this section, we show how to assess the algorithms implemented in the descriptive modeling section using a common graphical tool: the ROC curve (*see* **Note 10**).

3.3.1 ROC Curves

An interaction predicted by the algorithm is considered as a true positive (TP) or as a false positive (FP) depending on the presence or not of the corresponding interaction in the underlying "true" network, respectively. Analogously, the prediction of the absence of an interaction is considered as a true negative (TN) or a false negative (FN) depending on whether the corresponding edge is present or not in the underlying true network, respectively. Since GRN inference algorithms use a threshold value in order to define

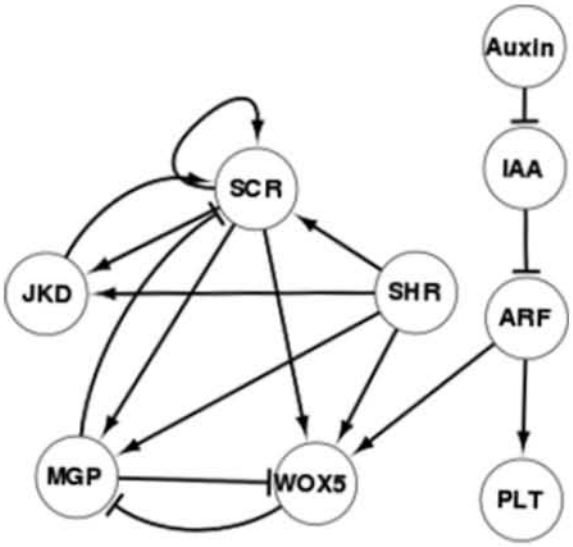


Fig. 4 "Real" root SCN GRN. The graph shows one of the single-cell Root SCN GRNs proposed in Azpeitia et al. [32]. The GRN is based on the experimental evidence summarized in Table 3, and it represents graphically the information encoded in the logical statements shown in Table 4

which edges are not included in the final network, the previous values (TP, FP, TN, and FN) can be calculated for each threshold value.

Using these definitions, two performance metrics can be calculated: the false positive rate, defined as FPR = FP/(TN + FP), and the true positive rate (sensitivity), TPR = TP/(TP + FN). The ROC curve is a commonly used graphical analysis in which the TPR (true positive rate) vs. FPR (false positive rate) are plotted for an inference algorithm as the threshold value is varied. A perfect inference algorithm would yield a point in the upper left corner of the ROC space, representing 100 % TPR (all true positives are found) and 0 % FPR (no false positives are found). Accordingly, points above the diagonal line indicate good inference results, while points below the line indicate wrong results.

A generic protocol to measure GRN Inference performance by means of a ROC curve analysis would be as follows:

1. Represent the inferred n genes network as an $n \times n$ adjacency matrix, where the cell i,j contains the value of similarity metric (PPC or MI) between the expression profiles of the genes i and j: both *cor* and *minet* functions return such a matrix (*see* Subheading 3.1).
2. Define an adjacency matrix for the "real" interactions, where the cell i,j contains 1(0) indicating the presence (absence) of experimentally supported interaction.
3. Use the function *validate()*, which takes as arguments the inferred and the real networks (in matrix form) and calculates the metrics TP, FP, TN, and FN (*see* Subheading 3.3.1) for different threshold values.
4. Measure the accuracy of each algorithm by calculating the area under the ROC curve using the function *auc.roc* of the package *minet*.

We applied the previous protocol to compare each of the inferred networks with the "real" experimentally supported network. Figure 5 shows the ROC curves for the four comparisons. The methods PPC and MRNET show a better performance, given that their curves (points) are closer to the top-left corner (perfect inference) than those of other methods. Table 6 shows the calculated AUC values. Interestingly, the simple PPC-based inference showed the highest accuracy, while the method ARACNE showed the lowest (*see* **Note 11**). Overall, the inference method shows a good performance (AUC > 8.3), with the exception of ARACNE. This suggests that inferred interactions from curated expression data set as the one assembled in [45] provide important background information consistent with experimentally supported functional relationships, at least for the module analyzed here.

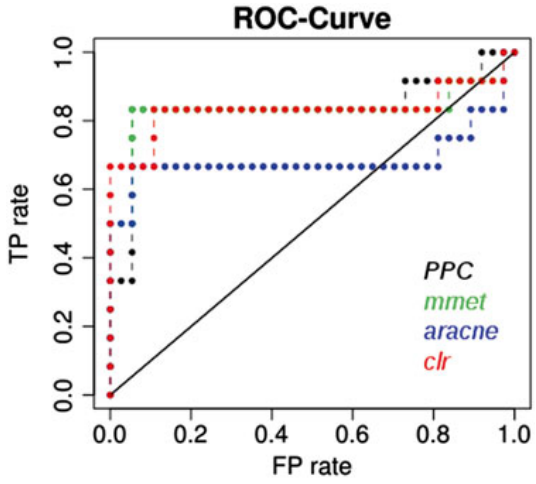


Fig. 5 ROC Curves for inference algorithms. The *graph* shows a comparison of the performance of each of the inference algorithms used herein. For each of the four algorithms, a ROC curve is plotted. Most of the points appear above the *diagonal line* indicating a general good inference performance. The curves that reach a higher TP rate while having low or null FP rate outperform the other. In this case: *clr*, *mrnet*, and *PPC* outperform *aracne*

Table 6
Area under the (ROC) curve (AUC) values

	PPC	CLR	ARACNE	MRNET
AUC	0.8355856	0.8333333	0.6869369	0.8310811

4 Notes

1. Correlation does not imply causation [61]. If two variables, A and B, are correlated with high statistical significance, it does not necessarily imply that A causes B (nor that B causes A).
2. Dataset selection is an important part in inference approaches. Finding or not interactions among variables directly depends on the statistical properties of the data. Depending on the goals of the study, one could choose to integrate a comprehensive large and heterogeneous dataset [46], or a smaller one based on certain selection criteria [45]. The results will likely vary depending on the dataset, even when using the exact same inference algorithm. The same is true for the performance of the different algorithms (*see* below).
3. Importantly, an edge in an inferred co-expression network does not imply a physical interaction or a direct regulatory influence. It is assumed that genes that are co-expressed across conditions are likely to share a common function, or to be

involved in similar biological processes [62]. This *functional* relationship does not imply a direct functional dependence between the corresponding molecules.

4. At first sight, from a mechanistic point of view, the entire notion of validating or invalidating models may seem misguided [17]. Models are valuable in science not because they can be validated, but because they can be useful for improving our understanding of a given observed phenomenon. Models may be found inconsistent with a set of data, but that does not necessarily rob them of their utility. The consequences of a specific set of assumptions included as underlying processes in a mechanistic model do not depend on the available experimental data, nor on a validation process. Thus, a mechanistic model is always a well-suited tool to address questions regarding such assumptions [63].
5. In the case of incomplete or uncertain prior knowledge about the system being modeled, a single model may be less useful than a set of models representing different hypotheses. Instead of having to decide if a specific model fits the data, which is hard and subjective, one can test which model fits the data best, which is easier and more objective [22]. In this way, putative interaction of functional relationships between genes can be postulated as hypotheses in the form of different GRN models. Each model can be tested against the observations (e.g., expected expression profiles) and in this way address which set of hypotheses fits better.
6. A link is established by an edge between two genes, represented by nodes, if the PCC value is higher or equal to an arbitrary cutoff that can be adjusted depending on the dataset used. In the present case, we chose the greatest value producing a fully connected network (a network where all the nodes have at least one edge). The chosen value was 0.3. In this case, such a small value is associated with the fact of having a fairly homogeneous dataset: only samples from a single tissue (root) and under wild-type conditions. Even in this case, the PCC-based inference showed good performance (*see* Subheading 3.3).
7. Unlike PPC, MI is not restricted to the identification of linear relations between the random variables, and is used as an approach to eliminate the majority of indirect interactions inferred by co-expression methods [47, 55].
8. A perfect coincidence would suggest that a sufficient set of molecular components (nodes) and a fairly correct set of interactions have been considered in the postulated GRN model. If this is not the case, additional components and interactions can be incorporated or postulated, or the Boolean functions can be

modified. This allows to refine interpretations of experimental data or to postulate novel interactions to be tested experimentally in the future. In any case, the process can be repeated several times based on the dynamical behavior of the modified versions of the GRN under study until a regulatory module is postulated. Such module can include some novel hypothetical interactions or components, integrate available experimental data, and identify possible experimental contradictions or gaps.

9. There are several free software packages for the dynamic analysis of Boolean GRNs, including: ANTELOPE [64], GINSIM [65], BoolNet [60], GNbox [66], GNA [67], and BioCham [68].
10. The performance of the inference algorithms heavily relies on the dataset used. There is no best algorithm for all cases. We showed that the simplest, most criticized algorithm (PPC-based inference) showed the best performance in the case analyzed here.
11. There are other tools to test the performance of inference algorithms. ROC curves can present an overly optimistic view of an algorithm's performance if there is a large skew in the types of interactions present in the true network (true and false interactions). This situation is common in GRN network inference because of sparseness. To tackle this problem, precision–recall (PR) curves can be used (*see* ref. 47).

Acknowledgments

J.D.V acknowledges the support of CONACYT and the Centre for Genomic Regulation (CRG), Barcelona, Spain; while spending a research visit in the lab of Stephan Ossowski. This chapter constitutes a partial fulfillment of the graduate program Doctorado en Ciencias Biomédicas of the Universidad Nacional Autónoma de México, UNAM in which J.D.V. developed this project. This work was supported by grants CONACYT 180098, 180380, 167705, 152649 and UNAM-DGAPA-PAPIIT: IN203113, IN 203214, IN203814, UC Mexus ECO-IE415. The authors acknowledge logistical and administrative help of Diana Romo.

References

1. Forgacs G, Newman SA (2005) Biological physics of the developing embryo. Cambridge University Press, Cambridge
2. Alvarez-Buylla ER, Benítez M, Dávila EB et al (2007) Gene regulatory network models for plant development. Curr Opin Plant Biol 10(1):83–91
3. Huang S, Kauffman S (2009) Complex gene regulatory networks—from structure to biological observables: cell fate determination. In: Meyers RA (ed) Encyclopedia of complexity and systems science. Springer, Heidelberg, pp 1180–1213
4. Alvarez-Buylla ER, Azpeitia E, Barrio R, Benítez M, Padilla-Longoria P (2010) From ABC genes to regulatory networks, epigenetic landscapes and flower morphogenesis: making

biological sense of theoretical approaches. Semin Cell Dev Biol 21(1):108–117
5. Kaneko K (2006) Life: an introduction to complex systems biology. Springer, New York
6. Azpeitia E, Alvarez-Buylla ER (2012) A complex systems approach to Arabidopsis root stem-cell niche developmental mechanisms: from molecules, to networks, to morphogenesis. Plant Mol Biol 80(4–5):351–363
7. Azpeitia E, Davila-Velderrain J, Villarreal C et al (2014) Gene regulatory network models for floral organ determination. In: Riechmann JL, Wellmer F (eds) Flower development. Springer, New York, pp 441–469
8. Barrio RÁ, Hernández-Machado A, Varea C, Romero-Arias JR, Alvarez-Buylla E (2010) Flower development as an interplay between dynamical physical fields and genetic networks. PLoS One 5(10):e13523
9. Barrio RÁ, Romero-Arias JR, Noguez MA et al (2013) Cell patterns emerge from coupled chemical and physical fields with cell proliferation dynamics: the *Arabidopsis thaliana* root as a study system. PLoS Comput Biol 9(5):e1003026
10. Proost S, Van Bel M, Sterck L et al (2009) PLAZA: a comparative genomics resource to study gene and genome evolution in plants. Plant Cell 21(12):3718–3731
11. Weigel D, Mott R (2009) The 1001 genomes project for *Arabidopsis thaliana*. Genome Biol 10(5):107
12. Hawkins RD, Hon GC, Ren B (2010) Next-generation genomics: an integrative approach. Nat Rev Genet 11(7):476–486
13. Lamesch P, Berardini TZ, Li D et al (2012) The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. Nucleic Acids Res 40(D1):D1202–D1210
14. Haughn GW, Somerville CR (1988) Genetic control of morphogenesis in Arabidopsis. Dev Genet 9(2):73–89
15. Rowan BA, Weigel D, Koenig D (2011) Developmental genetics and new sequencing technologies: the rise of nonmodel organisms. Dev Cell 21(1):65–76
16. Bowman JL, Smyth DR, Meyerowitz EM (2012) The ABC model of flower development: then and now. Development 139(22): 4095–4098
17. Lander AD (2010) The edges of understanding. BMC Biol 8(1):40
18. Yaffe MB (2013) The scientific drunk and the lamppost: massive sequencing efforts in cancer discovery and treatment. Sci Signal 6(269):pe13
19. Lee WP, Tzou WS (2009) Computational methods for discovering gene networks from expression data. Brief Bioinform 10(4):408–423
20. De Smet R, Marchal K (2010) Advantages and limitations of current network inference methods. Nat Rev Microbiol 8(10):717–729
21. Villaverde AF, Banga JR (2014) Reverse engineering and identification in systems biology: strategies, perspectives and challenges. J R Soc Interface 11(91):20130505
22. Ellner SP, Guckenheimer J (2011) Dynamic models in biology. Princeton University Press, Princeton, NJ
23. Kell DB, Oliver SG (2004) Here is the evidence, now what is the hypothesis? The complementary roles of inductive and hypothesis-driven science in the post-genomic era. Bioessays 26(1):99–105
24. Dehmer M, Emmert-Streib F, Graber A et al (2011) Applied statistics for network biology: methods in systems biology. Wiley, New York
25. Hartwell LH, Hopfield JJ, Leibler S (1999) From molecular to modular cell biology. Nature 402:C47–C52
26. Kashtan N, Alon U (2005) Spontaneous evolution of modularity and network motifs. Proc Natl Acad Sci U S A 102(39):13773–13778
27. Espinosa-Soto C, Wagner A (2010) Specialization can drive the evolution of modularity. PLoS Comput Biol 6(3):e1000719
28. Mitra K, Carvunis AR, Ramesh SK et al (2013) Integrative approaches for finding modular structure in biological networks. Nat Rev Genet 14(10):719–732
29. Mendoza L, Alvarez-Buylla ER (1998) Dynamics of the genetic regulatory network for *Arabidopsis thaliana* flower morphogenesis. J Theor Biol 193(2):307–319. doi:10.1006/jtbi.1998.0701
30. Espinosa-Soto C, Padilla-Longoria P, Alvarez-Buylla ER (2004) A gene regulatory network model for cell-fate determination during *Arabidopsis thaliana* flower development that is robust and recovers experimental gene expression profiles. Plant Cell 16:2923–2939
31. Albert R, Othmer HG (2003) The topology of the regulatory interactions predicts the expression pattern of the segment polarity genes in *Drosophila melanogaster*. J Theor Biol 223(1):1–18
32. Azpeitia E, Benítez M, Vega I, Villarreal C, Alvarez-Buylla ER (2010) Single-cell and coupled GRN models of cell patterning in the *Arabidopsis thaliana* root stem cell niche. BMC Syst Biol 4:134

33. Azpeitia E, Weinstein N, Benítez M et al (2013) Finding missing interactions of the *Arabidopsis thaliana* root stem cell niche gene regulatory network. Front Plant Sci 4:110
34. Caragea D, Welch SM, Hsu WH (2010) Handbook of research on computational methodologies in gene regulatory networks. Medical Information Science Reference, Hershey, PA
35. Wang R, Li C, Aihara K (2010) Modeling biomolecular networks in cells. Springer, New York
36. Lingeman JM, Shasha D (2012) Network inference in molecular biology. Springer, New York
37. Friedel S, Usadel B, Von Wirén N et al (2012) Reverse engineering: a key component of systems biology to unravel global abiotic stress cross-talk. Front Plant Sci 3:294
38. Usadel B, Fernie AR (2013) The plant transcriptome—from integrating observations to models. Front Plant Sci 4:48
39. Jaeger J, Sharpe J (2014) On the concept of mechanism in development. In: Minelli A, Pradeu T (eds) Towards a theory of development. Oxford University Press, Oxford, p 56
40. Hua F, Hautaniemi S, Yokoo R et al (2006) Integrated mechanistic and data-driven modelling for multivariate analysis of signalling pathways. J R Soc Interface 3(9):515–526
41. McGeachie MJ, Chang HH, Weiss ST (2014) CGBayesNets: conditional Gaussian Bayesian network learning and inference with mixed discrete and continuous data. PLoS Comput Biol 10(6):e1003676
42. Crombach A, Wotton KR, Cicin-Sain D et al (2012) Efficient reverse-engineering of a developmental gene regulatory network. PLoS Comput Biol 8(7):e1002589
43. Mao L, Van Hemert JL, Dash S et al (2009) Arabidopsis gene co-expression network and its functional modules. BMC Bioinformatics 10(1):346
44. Feltus FA, Ficklin SP, Gibson SM et al (2013) Maximizing capture of gene co-expression relationships through pre-clustering of input expression samples: an Arabidopsis case study. BMC Syst Biol 7(1):44
45. Montes RA, Coello G, González-Aguilera KL et al (2014) ARACNe-based inference, using curated microarray data, of *Arabidopsis thaliana* root transcriptional regulatory networks. BMC Plant Biol 14(1):97
46. Netotea S, Sundell D, Street NR et al (2014) ComPlEx: conservation and divergence of co-expression networks in *A. thaliana*, Populus and *O. sativa*. BMC Genomics 15(1):106
47. Meyer PE, Lafitte F, Bontempi G (2008) minet: AR/Bioconductor package for inferring large transcriptional networks using mutual information. BMC Bioinformatics 9(1):461
48. Hansen KD, Gentry J, Long L et al (2009) Rgraphviz: provides plotting capabilities for R graph objects. R package version 2.8.1. 2009.
49. Usadel B, Obayashi T, Mutwil M et al (2009) Co-expression tools for plant biology: opportunities for hypothesis generation and caveats. Plant Cell Environ 32(12):1633–1651
50. Cho DY, Kim YA, Przytycka TM (2012) Network biology approach to complex diseases. PLoS Comput Biol 8(12):e1002820
51. Cramer GR, Urano K, Delrot S et al (2011) Effects of abiotic stress on plants: a systems biology perspective. BMC Plant Biol 11(1):163
52. Ficklin SP, Feltus FA (2011) Gene coexpression network alignment and conservation of gene modules between two grass species: maize and rice. Plant Physiol 156(3):1244–1256
53. Hernández-Lemus E, Velázquez-Fernández D, Estrada-Gil JK et al (2009) Information theoretical methods to deconvolute genetic regulatory networks applied to thyroid neoplasms. Phys Stat Mech Appl 388(24):5057–5069
54. Meyer PE, Olsen C, Bontempi G (2011) Transcriptional network inference based on information theory. In: Dehmer M, Emmert-Streib F et al (eds) Applied statistics for network biology: methods in systems biology. Weinheim, Wiley-Blackwell, pp 67–89
55. Margolin AA, Nemenman I, Basso K et al (2006) ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. BMC Bioinformatics 7(Suppl 1):S7
56. Sales G, Romualdi C (2011) parmigene—a parallel R package for mutual information estimation and gene network reconstruction. Bioinformatics 27(13):1876–1877
57. Kauffman S (1969) Homeostasis and differentiation in random genetic control networks. Nature 224:177–178
58. Albert I, Thakar J, Li S, Zhang R, Albert R (2008) Boolean network simulations for life scientists. Source Code Biol Med 3:16
59. Davila-Velderrain J, Martinez-Garcia JC, Alvarez-Buylla ER (2014) Epigenetic landscape models: the post-genomic era. *bioRxiv*
60. Müssel C, Hopfensitz M, Kestler HA (2010) BoolNet - an R package for generation, reconstruction and analysis of Boolean networks. Bioinformatics 26(10):1378–1380
61. Huang S (2014) When correlation and causation coincide. Bioessays 36(1):1–2
62. Lehner B, Lee I (2008) Network-guided genetic screening: building, testing and using

gene networks to predict gene function. Brief Funct Genomic Proteomic 7(3):217–227

63. Gershenfeld N (1999) The nature of mathematical modeling. Cambridge University Press, Cambridge
64. Arellano G, Argil J, Azpeitia E et al (2011) "Antelope": a hybrid-logic model checker for branching-time Boolean GRN analysis. BMC Bioinformatics 12:490
65. Naldi A, Berenguier D, Fauré A et al (2009) Logical modeling of regulatory networks with ginsim 2.3. Biosystems 97(2):134–139
66. Corblin F, Fanchon E, Trilling L (2010) Applications of a formal approach to decipher discrete genetic networks. BMC Bioinformatics 11(1):385
67. De Jong H, Geiselmann J, Hernandez C et al (2003) Genetic network analyzer: qualitative simulation of genetic regulatory networks. Bioinformatics 19(3):336–344
68. Calzone L, Fages F, Soliman S (2006) Biocham: an environment for modeling biological systems and formalizing experimental knowledge. Bioinformatics 22(14):1805–1807

frontiers in Genetics

REVIEW
published: 23 April 2015
doi: 10.3389/fgene.2015.00160

Modeling the epigenetic attractors landscape: toward a post-genomic mechanistic understanding of development

Jose Davila-Velderrain[1,2], *Juan C. Martinez-Garcia*[3] *and Elena R. Alvarez-Buylla*[1,2]*

[1] Departamento de Ecología Funcional, Instituto de Ecología, Universidad Nacional Autónoma de México, Mexico City, Mexico, [2] Centro de Ciencias de la Complejidad (C3), Universidad Nacional Autónoma de México, Mexico City, Mexico, [3] Departamento de Control Automático, Cinvestav-Instituto Politécnico Nacional, Mexico City, Mexico

OPEN ACCESS

Edited by:
Moisés Santillán, Centro de Investigación y Estudios Avanzados del IPN, Mexico

Reviewed by:
David McMillen, University of Toronto Mississauga, Canada
Enrique Hernandez-Lemus, National Institute of Genomic Medicine, Mexico
Edgardo Ugalde, Universidad Autónoma de San Luis Potosí, Mexico

***Correspondence:**
Elena R. Alvarez-Buylla, Instituto de Ecología, Universidad Nacional Autónoma de México, 3er Circuito Exterior, Junto a Jardín Botánico, Mexico City, D.F. 04510, México
eabuylla@gmail.com

Specialty section:
This article was submitted to Systems Biology, a section of the journal Frontiers in Genetics

Received: *01 March 2015*
Accepted: *08 April 2015*
Published: *23 April 2015*

Citation:
Davila-Velderrain J, Martinez-Garcia JC and Alvarez-Buylla ER (2015) Modeling the epigenetic attractors landscape: toward a post-genomic mechanistic understanding of development. Front. Genet. 6:160. doi: 10.3389/fgene.2015.00160

Robust temporal and spatial patterns of cell types emerge in the course of normal development in multicellular organisms. The onset of degenerative diseases may result from altered cell fate decisions that give rise to pathological phenotypes. Complex networks of genetic and non-genetic components underlie such normal and altered morphogenetic patterns. Here we focus on the networks of regulatory interactions involved in cell-fate decisions. Such networks modeled as dynamical non-linear systems attain particular stable configurations on gene activity that have been interpreted as cell-fate states. The network structure also restricts the most probable transition patterns among such states. The so-called Epigenetic Landscape (EL), originally proposed by C. H. Waddington, was an early attempt to conceptually explain the emergence of developmental choices as the result of intrinsic constraints (regulatory interactions) shaped during evolution. Thanks to the wealth of molecular genetic and genomic studies, we are now able to postulate gene regulatory networks (GRN) grounded on experimental data, and to derive EL models for specific cases. This, in turn, has motivated several mathematical and computational modeling approaches inspired by the EL concept, that may be useful tools to understand and predict cell-fate decisions and emerging patterns. In order to distinguish between the classical metaphorical EL proposal of Waddington, we refer to the *Epigenetic Attractors Landscape* (EAL), a proposal that is formally framed in the context of GRNs and dynamical systems theory. In this review we discuss recent EAL modeling strategies, their conceptual basis and their application in studying the emergence of both normal and pathological developmental processes. In addition, we discuss how model predictions can shed light into rational strategies for cell fate regulation, and we point to challenges ahead.

Keywords: GRN, epigenetic landscape, attractors, cell-fate, morphogenesis, stem-cells, cancer

1. Introduction

The progressive loss of potency from pluripotent stem cells to mature, differentiated cells, as well as the reproducible emergence of spatiotemporal patterns through the course of development has

been always perceived as strong evidence of the robustness and *deterministic* nature of development. The explanation of such a robust process has puzzled researchers for many years. For a long time, although not always stated explicitly, the prevailing paradigm in developmental biology was supported on two fundamental paradigms: (1) a mature cell, once established, displays an essentially irreversible phenotype; and (2) the developmental process is controlled by a "program" as a genomic blueprint following a simplistic linear scheme of causation in an essentially deterministic fashion. Experimental and theoretical studies in the last decade have challenged these assumptions. It has been shown that a differentiated state of a given cell is not irreversible as previously thought, and that in fact, it is possible to reprogram differentiated cells into pluripotent states with a plethora of protocols in plants and animals (Grafi, 2004; Takahashi and Yamanaka, 2006; Takahashi et al., 2007; González et al., 2011).

Overall, a growing body of empirical evidence now supports intrinsic physical processes as a fundamental source of order instead of deterministic pre-programmed rules (Huang, 2009; Mammoto and Ingber, 2010). Although these observations have just recently shift the focus of study in developmental biology and biomedical research, the new evidence is in line with the proposals that early theoretical biologists posited decades ago (see, for example Waddington, 1957; Goodwin, 1963; Kauffman, 1969, 1993; Goodwin, 2001). C. H. Waddington was one of the first to point out that the physical implementation of the information coded in the genes and their interactions imposes developmental constraints while forming an organism. Waddington's heuristic model of the epigenetic landscape (EL) was a visionary attempt to consolidate these ideas in a conceptual framework that enables the discussion of the relationship between genetics, development, and evolution in an intuitive manner. Waddington's proposal was inspired in a formal dynamic systems approach, nonetheless (Waddington, 1957; Gilbert, 1991; Slack, 2002).

Nowadays in the data-rich, post-genomic era the EL has been consolidated as a useful conceptual model for the discussion of the mechanistic basis underlying cellular differentiation—particularly trans-differentiation and reprogramming events (Alvarez-Buylla et al., 2008; Enver et al., 2009; Fagan, 2012; Ladewig et al., 2013). This field has become particularly active due to its potential medical applications using stem cells systems biology as a means for discovering efficient reprogramming or therapeutic strategies by combining mathematical and computational modeling with experimental techniques (MacArthur et al., 2008, 2009; Roeder and Radtke, 2009; Huang, 2011; Zhou and Huang, 2011). Recently, though, numerous critiques to Waddington's original model have been presented in light of the dynamical plasticity of differentiated cells (see, for example Balázsi et al., 2011; Ferrell, 2012; Furusawa and Kaneko, 2012; Garcia-Ojalvo and Arias, 2012; Sieweke, 2015). In this review, we claim that the formalization of the EL in the context of the study of the dynamical properties of GRNs enables a formal framework which provides the necessary flexibility for a model to be both: (1) consistent with the observed inherent plasticity of developing cells and (2) formally derived from the uncovered regulatory underpinnings of cell-fate regulation. It is thus important to note that this GRN associated EL model is not to be confused with the literal, metaphorical model presented by Waddington, which some authors have associated only to the static diagrammatic proposal originally put forward (West-Eberhard, 2003). In order to highlight such distinction, here we will refer to the EL model associated with the dynamics of GRNs as the *epigenetic attractors landscape* (EAL).

The conceptual distinction between the classical EL and the EAL proposed here, as well as its relevance as a consistent model for the prevailing theories of differentiation is going to be exposed by the authors elsewhere. In this contribution we instead focus on the mathematical approaches which have been developed to derive an EAL as an extension of the conventional dynamical analyses of experimentally grounded GRN models. Importantly, we deliberately use the generic term EAL to refer to a group of dynamical models which are quite diverse in mathematical properties and structure, however we do so for phenomenological reasons: all the approaches try to formally tackle the phenomenon of cellular differentiation taking the classical EL model as a conceptual basis. Given the current relevance of such a modeling exercise applied to molecular networks involved on processes such as stem cell differentiation (Li and Wang, 2013), tissue morphogenesis (Alvarez-Buylla et al., 2010), and carcinogenesis (Choi et al., 2012; Wang et al., 2014); and the fact that different approaches have been proposed in order to reach similar goals (Huang, 2009, 2012; Zhou et al., 2012), we hope that the present integrative review may prove useful for a wide range of biological applications. Our main objective is 2-fold: (1) to help different research groups attempting to formalize the EAL to reduce the gap existing between current different approaches and (2) to contribute to shape a common and formal discussion ground on EAL models among experimentalists and theoretical biologists. Accordingly, we have decided to favor conceptual clarity over technicalities through the text, and to point to original references where more detail is available if necessary. We apologize for the theoretically oriented reader for the lack of mathematical formality.

1.1. The Dynamical-Systems View of Cell Biology

The modern picture of the EL is framed in the context of GRN dynamics (Kauffman, 1969; Mendoza and Alvarez-Buylla, 1998; Huang, 2012), and its theoretical basis is a dynamical-systems perspective. From here on we will refer to this view of the EL model as the EAL. Under dynamical-systems framework a cell is considered a dynamical system, assuming that its state at a certain time can be described by a set of time-dependent variables. As a first approximation, it is commonly assumed that the amount of the different proteins within the cell or, for practical reasons, the levels of gene expression (i.e., expression profiles) are sufficient to describe such state (Huang, 2013). Thus, the expression profile is conventionally taken as the set of variables representing the state of the cell; each gene in the cell's GRN representing one variable (see **Figure 1**). Mathematically, the set of variables is represented as a state vector given by $\mathbf{x}(t) = [x_1(t), x_2(t)..., x_n(t)]$ for a GRN with n genes. Given such specification, it is useful to imagine an abstract space termed the *state space* of the system. In the context of GRNs the state space comprises all the theoretically possible states a cell can exhibit; each point in this abstract space represents one particular expression profile (**Figure 1B**).

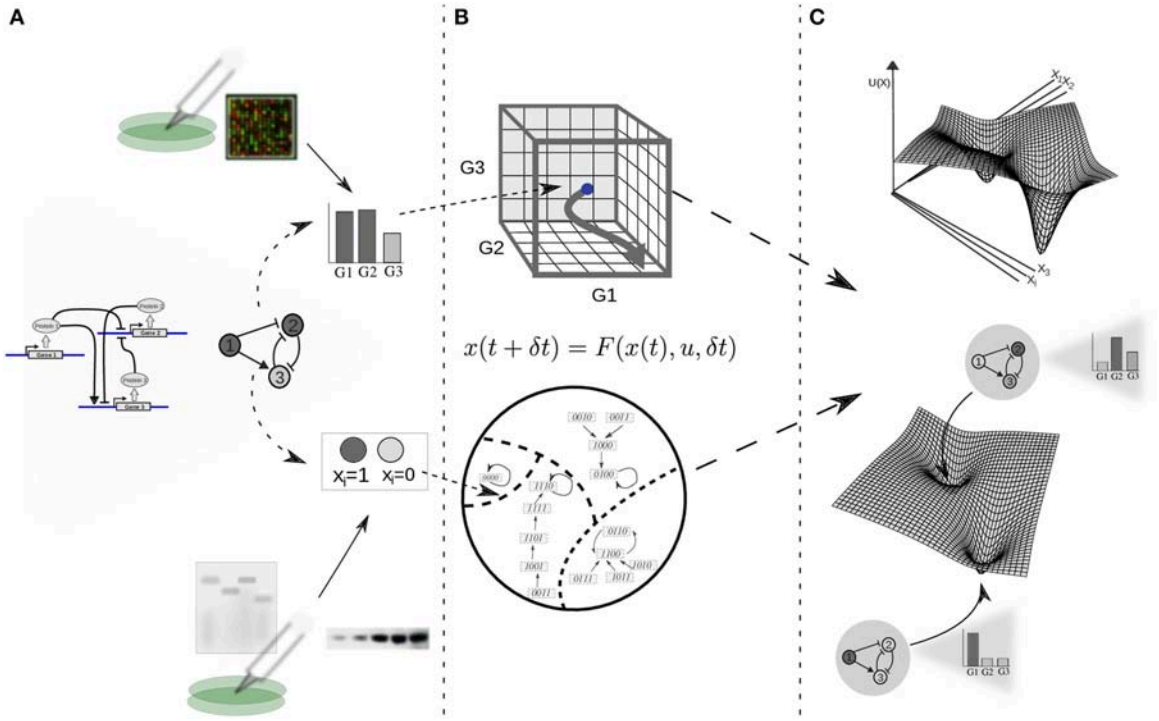


FIGURE 1 | From experimental data on gene function and interactions to a dynamic gene regulatory network and epigenetic landscape model. (A) The architecture of a GRN is proposed given available experimental molecular data; the state of the network is specified as a gene expression profile or gene on/off (1/0) configuration for the case of continuous or discrete state models, respectively. Boolean or differential equations are used, respectively. **(B)** The complete set of possible states define a continuous (above) or discrete (below) state space, where each state corresponds to a point; changes in gene expression during developmental dynamics manifest as trajectories in this abstract space (here depicted as arrows). **(C)** In an intuitive characterization of an EL, an "elevation" value $U(x)$ is associated to each network state x. The association of "elevation" values to network states, or more generally, the quantitative characterization of their relative stability is the ultimate goal of EL modeling efforts. For illustrative purposes, the EL is depicted here as a hypothetical low-dimensional projection.

Furthermore, it is assumed that the cell state at a certain time and the cell state at a later time are connected by a state trajectory in a causal way.

Mathematically, the current cell state is a function or a more general mapping of the initial state and certain additional parameters. The connection between cell states can be formally expressed by a dynamical equation,

$$\mathrm{x}(t + \delta t) = \mathbf{F}(\mathbf{x}(t), \mathbf{u}, \delta t), \tag{1}$$

where **F** represents the map that connects one state with the immediately previous sate (**F** is also known as the *transition map*), $\mathbf{x}(t)$ denotes the state at a certain time t, and **u** stands for the vector of additional parameters. Both the time increments δt and the state variables $x_i(t)$ can be either continuous or discrete, depending on the chosen mathematical formalism. Within the cell, the map **F** is implemented by the architecture of the GRN, which specifies both the topology of the network and the nature and form of the corresponding gene regulations (Huang, 2009). Because of globally conditioned gene behavior due to mutual gene regulatory interactions, through the causal connections between cell states, the GRN imposes dynamical constraints and limits the permissible behavior of the cell. Of special interest are the transient and emergent stable configurations that the cell may attain as a result. The existence of the dynamical map **F** expresses the causality of the cellular developmental process and the mechanistic character of GRN dynamical models.

One of the most salient and impressive features of GRNs is the existence of a small number of stationary or quasi-stationary gene configurations within the state space (Kauffman, 1969). Given a specific GRN, a set of cell states satisfy the constraints imposed by the GRN; that is, each of these cell states is connected to itself by the map **F** (i.e., $\mathbf{x}^* = \mathbf{F}(\mathbf{x}^*, \mathbf{u})$). When these steady states ($\mathbf{x}^*$) are also resilient to perturbations, that is, if they return back to the steady state after being kicked away by state variations either of intrinsic or external origin, we refer to them as *attractors*. In the case of quasi-stationary states, if a set comprised of several individual states repeats in a cyclic manner it corresponds to a cyclic attractor. All other states are either unstable or form part of transitory trajectories channeled toward one of these attractor states. The theory posits that attractor states correspond to the observable robust cell phenotypes, cell types, or cellular processes; and that these emerge as a natural consequence of the dynamical constraints imposed by the underlying GRN (Huang and Kauffman, 2009; Huang, 2013). For a more formal definition of attractors in dynamical systems theory see (Fuchs, 2013a).

1.2. Extending GRN to EAL Models

The postulation of experimentally grounded GRN dynamical models, their qualitative analysis and dynamical characterization in terms of control parameters, and the validation of predicted attractors against experimental observations has become a well-established framework for the study of developmental dynamics in systems biology—see, for example: (Mendoza and Alvarez-Buylla, 1998; Von Dassow et al., 2000; Albert and Othmer, 2003; Espinosa-Soto et al., 2004; Huang et al., 2007; Graham et al., 2010; Sciammas et al., 2011; Hong et al., 2012; Jaeger and Crombach, 2012; Azpeitia et al., 2014). The qualitative analysis of the dynamics of GRN models is well-suited for the study of the specification of cell fates as a result of the constrains imposed by the associated GRN. This conventional analysis includes the identification and local characterization of attractor states, and the comparison of these predicted cell-type configurations with the ones that are actually observed in the corresponding biological system (**Figures 1A,B**).

If one is interested in studying the potential transition events among the already characterized stable cellular phenotypes, however, several difficulties arise. Standard analysis of dynamical systems, which focuses on the existence and local properties of a given attractor, fail to capture the main problem which is concerned with the relative properties of the different attractors (Zhou et al., 2012). In deterministic GRN models, given certain values for the related control parameters, the system under study always converges to a single attractor if initialized from the same state, and once it attains such steady-state it remains there indefinitely. In contrast, during a developmental process, cells change from one stable cell configuration to another one in specific temporal and spatial or morphogenic patterns. Additional formalisms are needed in order to explore questions regarding how cells in the course of differentiation transit among available given attractors, or the order in which the system converges to the different attractors given an initial condition; as well as to predict how these mechanisms can be altered by rational strategies.

1.2.1. EAL Modeling Goals

The need for extending GRN dynamical models beyond standard local analysis is related with the interest in addressing the following—and similar—questions. Conceptually, given an experimentally determined GRN, how can we explain and predict both specific "normal" and altered cellular differentiation events or morphogenic patterns? Is it possible to control the fate of differentiation events through well-defined stimuli? Can we deliberately cause altered morphogenic patterns by means of either genetic, physical, chemical or other type of environmental perturbations? Or formally, given a specific dynamical mapping $\mathbf{x}(t+\delta t) = \mathbf{F}(\mathbf{x}(t), \mathbf{u}, \delta t)$, and its associated state space, how can we study the conditions under which a transition event occurs among the attractor states $\mathbf{x}^*$? Is there a reproducible pattern of transitions? Can we alter the expected pattern through specific external control perturbations $\mathbf{u}$? To what extent are the observed robust and altered temporal or spatial morphogenetic patterns emergent consequences of the GRNs? The extension of GRN dynamical models and their analysis in order to address these and similar questions has shown to be a fruitful area of research in recent years (Han and Wang, 2007; Alvarez-Buylla et al., 2008; Wang et al., 2010b, 2014; Choi et al., 2012; Qiu et al., 2012; Villarreal et al., 2012; Zhou et al., 2012; Li and Wang, 2013; Zhu et al., 2015). The conceptual basis for most of these efforts is the EAL.

1.3. Deterministic EAL Models from Genetic Circuits

1.3.1. An Introductory Toy Model

A quite simple auto-activating single-gene circuit, a basic model of cell differentiation induction, is exposed in Ferrell (2012) as a conceptual tool to discuss some difficulties regarding Waddington's EL. In this work an EAL is mathematically described by a potential function. In dynamical systems theory, besides the state space approach explained briefly above, there is another way to visualize the dynamics of a system, but applicable only if the system is simple enough: the potential function (Strogatz, 2001; Fuchs, 2013a). The potential is a function $V(x)$ which (in one-dimensional systems) fulfills the relation given by:

$$\frac{dx}{dt} = f(x) = -\frac{dV(x)}{dx}, \tag{2}$$

i.e., $f(x)$ is the negative derivative of the potential, which can be found by direct integration:

$$V(x) = -\int f(x)dx. \tag{3}$$

Such a function defines an attractor landscape for the given dynamical system, and its plot graphically represents the dynamics of the system (**Figure 2**). Specifically, *minima* of the potential correspond to fixed-point attractors (e.g., cell types), and *maxima* correspond to unstable fixed-points. The motion, i.e., the state trajectories are given by the gradient lines (the lines of steepest descent of the potential). The trajectories are attracted by the minima of the potential. This corresponds to an intuitive, direct derivation of the EAL: a "height" value is associated to each of the points in the state-space in a way that those regions corresponding to attractors will have a lower value than that of the other transitory states (**Figure 2C**). Conceptually, the rolling ball of Waddington's EL will represent the state of a differentiating cell moving from higher to lower regions in state space. Thus, the calculated heights of the different attractors are expected to reflect their developmental potential in a hierarchical way: the lower height the lower potential for differentiation.

All one-dimensional systems have a potential function, but most two- or higher-dimensional systems do not (Fuchs, 2013a). This means that one could only apply this method if the cell is represented by a single-gene (single variable) circuit. Furthermore, note that here the EAL plays the role of a "toy" model useful in conceptual discussions, a role quite relevant (see Ferrell, 2012) but similar to that of the original metaphorical proposal of Waddington. In this review we devote more attention to the application of EAL models to real specific developmental

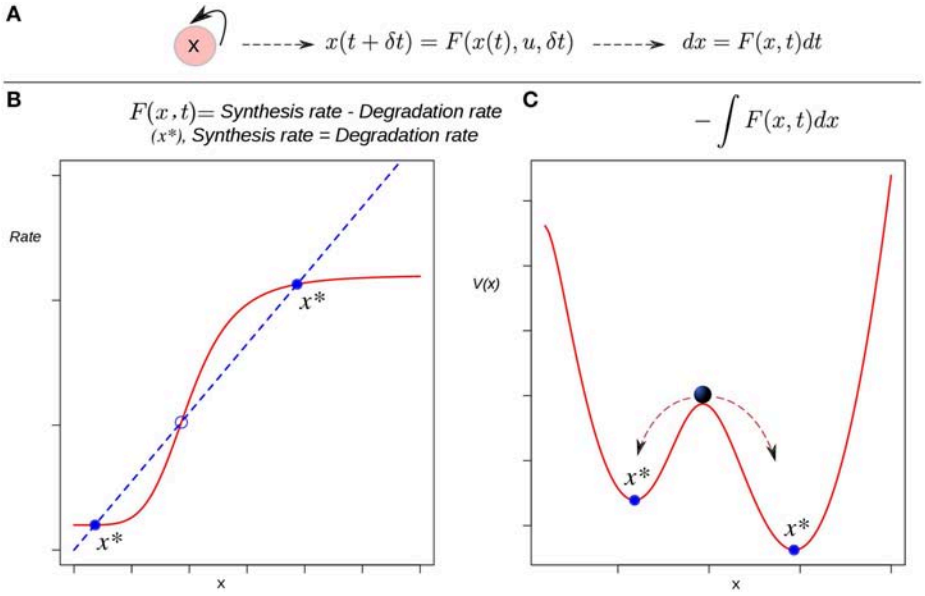


FIGURE 2 | The derivation of a potential function to visualize the epigenetic landscape and the dynamics of a system. (A) The causal connection between the state of the system (an auto-activating gene circuit) at a certain time and its state at a later time is modeled by a differential equation. **(B)** Attractors in rate-balance analysis. The blue dotted line and the red curve represent respectively a linear degradation rate, and a non-linear synthesis rate for the circuit's gene; the restrictions imposed by the circuit to the systems dynamics are met when both rates are balanced. The states that meet this balanced condition are stationary, and if stable (filled circles), are denoted as attractor states x^*. Circles represent stationary states. **(C)** The potential function. The attractor states x^* lie at the bottoms of valleys (minima). The trajectories starting from unstable, transitory states are attracted by the minima of the potential. The relative stability of the left (right) attractor with respect to the other is lower (higher) as quantified by the lower (higher) barrier height between them.

processes with explanatory and predictive purposes that generally involve n-dimensional GRN. Thus, a more "realistic" subnetwork model incorporating several transcription factors in a modular structure is necessary in such cases. The application of the integration-based potential function approach, however, cannot be applied to cases with a higher number of genes. Also, one should be cautious when postulating the existence of a potential for living systems in strict sense: a cell is an open non-equilibrium thermodynamical system, and its dynamics in general does not follow a gradient (since the transition rate between two given attractor states is not path-independent). For details, see (Zhou et al., 2012; Huang, 2013). For this reason authors use the term "quasi-potential" when speaking about cellular dynamics from a system-dynamics point of view (see below).

In the general case, the dynamics of continuous-time models of GRNs is given by more general types of autonomous differential equations (DEs). The time evolution of the cell state $\mathbf{x}(t)$ is commonly modeled by the system of DEs:

$$\frac{dx_i(t)}{dt} = F_i(x_1, x_2, ..., x_f, \mathbf{u}), \tag{4}$$

where $i = 1, 2, ..., n$ for a GRN of n genes. A dynamics defined by such a general DE is a special form of the map in Equation (1). In general, the functions $\mathbf{F}$ in the continuous-time model for cellular dynamics (Equation 4) are non-linear, and cannot be analytically integrated and derived from a gradient. Numerical approaches have been proposed to draw a deterministic "quasi-potential" for two-gene circuits (see, for example Bhattacharya et al., 2011). In what follows we focus on medium size GRN modules, where neither the direct integration nor the numerical deterministic approach are applicable. We start with the simplest models of GRN dynamics.

1.4. Stochastic EAL Models from Boolean GRNs

The first computational model envisioned for the simulation and analysis of the dynamic behavior of GRNs was the Boolean Network (BN) model (Kauffman, 1969, 1993). This model has been extended to model various developmental processes in the context of the EAL (Han and Wang, 2007; Alvarez-Buylla et al., 2008; Ding and Wang, 2011; Choi et al., 2012; Flöttmann et al., 2012). A BN models a dynamical system assuming both discrete time and discrete state. This is expressed formally with the mapping:

$$x_i(t+1) = F_i(x_1(t), x_2(t), ..., x_f(t)), \tag{5}$$

where the set of functions F_i are logical propositions expressing the relationship between the genes that share regulatory interactions with the gene i, and where the state variables $x_i(t)$ can take the discrete values 1 or 0 indicating whether the gene i is active or not at a certain time t, respectively. An experimentally grounded Boolean GRN model is then completely specified by the set of genes proposed to be involved in the process of interest and the associated set of logical functions derived from experimental data (Azpeitia et al., 2014). A dynamics defined by such a mapping is a special form of the map in Equation (1).

1.4.1. Attractor Transition Probability Approach to Explore the EAL

As stated above, in a deterministic framework, once a cell state corresponds to an attractor, it will remain there indefinitely. The set of conditions that lead to each attractor comprise the attracting *basin*. Under stochastic fluctuations, the borders of attractor regions in state space may be reached and may be crossed, leading to transitions from one attractor to another one (Ebeling and Feistel, 2011). Thus, the implementation of an stochastic dynamical model opens the opportunity to study signal-independent transitions among attractors. There are several approaches to include stochasticity in dynamical models. One approach is based on the idea of introducing transition probabilities. As discussed above, when studying cellular developmental dynamics, the transitions of interest are those among attractor states. Can these transitions be studied in terms of probabilities? Indeed, since Boolean GRN can be extended to include stochasticity and transition probabilities among attractors can then be estimated. Several ways to include stochasticity in a Boolean GRN model have been proposed (Garg et al., 2012). One way is the so-called stochasticity in nodes (SIN) model. Here, a constant probability of error ξ is introduced for the deterministic Boolean functions. In other words, at each time step, each gene "disobeys" its Boolean function with probability ξ. Formally:

$$\begin{aligned} P_{x_i(t+1)}[F_i(\mathbf{x}_{reg_i}(t))] &= 1-\xi, \\ P_{x_i(t+1)}[1-F_i(\mathbf{x}_{reg_i}(t))] &= \xi. \end{aligned} \tag{6}$$

The probability that the value of the now random variable $x_i(t+1)$ is determined or not by its associated logical function $F_i(\mathbf{x}_{reg_i}(t))$ is $1-\xi$ or ξ, respectively.

Alvarez-Buylla and collaborators used this extended BN model to explore the EAL associated with an experimentally grounded GRN (Alvarez-Buylla et al., 2008) (see below). In a BN model the set of possible states is finite. Specifically, due to its binary state character the state space of a Boolean GRN with n genes has a size of 2^n and is composed by the set of all possible binary vectors of length n (see **Figure 3A**). By simulating a stochastic one-step transition, according to the model in Equation (6) and the mapping in Equation (5), and starting from each of all the possible states in the system for a large number of times, it is possible to estimate the probability of transition from an attractor i to an attractor j as the frequency of times the states belonging to the basin of the attractor i are mapped into a state within the basin of the attractor j. For detail see (Azpeitia et al., 2014). In Alvarez-Buylla et al. (2008), the authors followed this simulation approach to estimate a transition probability matrix Π with components:

$$\pi_{ij} = P(A_{t+1} = j | A_t = i), \tag{7}$$

representing the probability that an attractor j is reached from an attractor i (**Figure 3B**). Once the set of attractors is known and the transition probability matrix is estimated, it is straightforward to implement a discrete time Markov chain model (DTMC) and obtain a dynamic equation for the probability distribution (for details, see Allen, 2010):

$$P_A(t+1) = \Pi P_A(t), \tag{8}$$

where $P_A(t)$ is the probability distribution over the attractors at time t, and Π is the transition probability matrix previously estimated. This equation can be iterated to simulate the temporal evolution of the probability distribution over the attractors starting from a biologically meaningful initial distribution. The extension of a Boolean GRN in order to apply this approach is quite simple and intuitive; however, there is a limitation that impedes its general applicability: as the size of the GRN grows, it becomes difficult to exhaustively characterize the attractor's landscape associated with the GRN in terms of the emergent attractors and its corresponding basins of attraction. If the dynamics of the Boolean GRN is not exhaustively characterized, the corresponding transition probabilities among attractors cannot be estimated using the proposed approach. Additionally, other implementations of stochasticity within BN models have been discussed (Garg et al., 2012). Additional examples should be worked out with such various approaches to test which is more practical and if all yield equivalent results.

1.4.2. Probabilistic Landscape (Quasi-potential) Approach

Han and Wang proposed a different approach in order to extend a BN model. Their goal was to first estimate the one-step transition probabilities among all the possible states in the state space and not just among given attractors (Han and Wang, 2007). For this, they implemented a variation of the BN that was previously proposed by Li and collaborators (Li et al., 2004) and which has been called the threshold network formalism (Thompson and Galitski, 2012). In this model, the structure of the network is formally represented with an adjacency matrix **C**, whose components c_{ij} indicating the nature and strength of the interaction from the gene j to gene i. The dynamic mapping for this BN model takes the form:

$$x_i(t+1) = \begin{cases} 1, & \sum_j c_{ij}x_j(t) + b_i > 0, \\ 0, & \sum_j c_{ij}x_j(t) + b_i < 0, \\ x_i(t), & \sum_j c_{ij}x_j(t) + b_i = 0, \end{cases} \tag{9}$$

where b_i is a parameter representing the ground state of the gene i: its state in the absence of regulation. The set of parameters (i.e., b_i and c_{ij}) can be chosen to force the dynamics of the BN to be consistent with those of a BN with a specific set of logical propositions (for details, see Supplementary Material in Choi et al., 2012). The mapping in Equation (9) can be conceptualized as follows: if the total input of a gene in the network is positive (activation), negative (repression) or zero; the future state of the gene will be active, inactive or unchanged from its previous state, respectively. Here, the total input of a gene is the sum of the previous states of the genes regulating it. The characterization of the

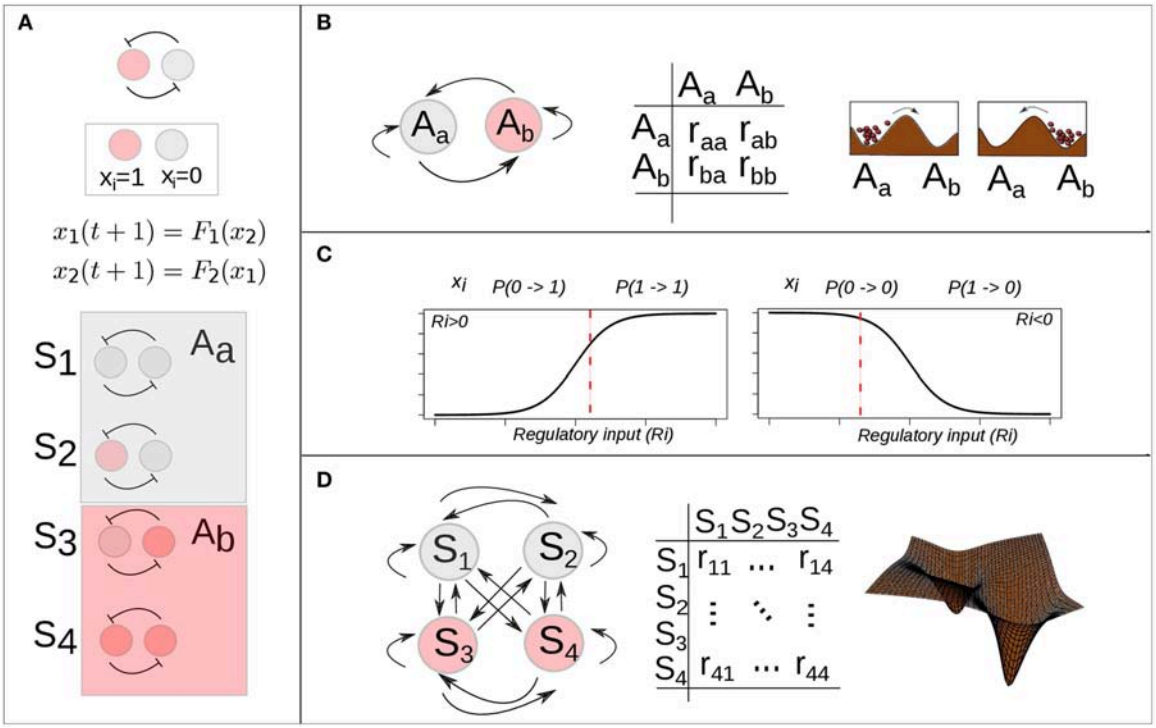


FIGURE 3 | Stochastic epigenetic landscape models from Boolean dynamics. (A) A simple mutual-inhibition circuit is modeled as a Boolean network: discrete temporal evolution and binary (0,1) state variable. The discrete state space corresponds to the set of binary vectors (here 4 possible states) and is partitioned by two basins of attraction. **(B)** There are $4(2^2)$ possible transitions among the two attractors. A 2×2 transition probability matrix specifies the probability of each possible transition. **(C)** Han and Wang proposed the use of a sigmoidal function of the total regulatory input (Ri) to calculate the probability of a one-step state transition of one gene *i* (Han and Wang, 2007). A specific value of the function (vertical red dotted line) gives the probability of the gene *i* becoming active (left) or inactive (right), given its regulatory input (Ri) in the current time. **(D)** There are $16(2^2 \times 2^2)$ possible transitions among the $4(2^2)$ possible states. A $2^2 \times 2^2$ transition probability matrix specifies the probability of each possible transition.

entire attractor's landscape can then be done through numerical iterations of this dynamical map as long as the network has a moderate size.

Han and Wang extended the deterministic BN model into a probabilistic framework by introducing a transition probability matrix. However, if the interest is focused on the computation of the probability of transition from one state to another state for each of the 2^n possible phenotypes in state space, then it is necessary to introduce a transition probability matrix with the probability of all possible transitions and not just among attractors. In order to make such computation feasible, Han and Wang introduced a simplification: they assumed that the one-step transition probability of one state to another can be expressed as the product of the probability of each gene in the network being activated or not, given the state of the network in the previous time (for details, see Han and Wang, 2007, and Supplementary Material in Choi et al., 2012). Formally:

$$\pi_{kj} = P(\mathbf{x}(t+1) = k | \mathbf{x}(t) = j) = \prod_{i=1}^{n} P(x_i(t+1) | \mathbf{x}(t) = j), \quad (10)$$

where j and k represent two different cell states and can take values from $[1, \ldots, 2^n]$; n is the number of genes in the network. The factorized transition probabilities are calculated by inserting a non-zero regulatory input ($\sum_{j=1}^{n} c_{ij} x_j(t) + b_i(t) \neq 0$) as the argument of a sigmoidal function whose range spans from 0 to 1, which is to say:

$$P(x_i(t+1) = 1 | \mathbf{x}(t) = j) = \frac{1}{2} \pm \frac{1}{2} \tanh \left[\mu \sum_{j=1}^{n} c_{ij} x_j + b_i \right]. \quad (11)$$

In the case of no input (i.e., $\sum_{j=1}^{n} c_{ij} x_j(t) + b_i = 0$) a small-valued parameter d is introduced:

$$P(x_i(t+1) = x_i(t) | \mathbf{x}(t) = j) = 1 - d.$$

Hence, in this approach, the probability that a gene *i* will be active (1) at a future time $t+1$ will be closer to one as long as its total input at the previous time t is high. Similarly, the probability of being inactive (0) at the future time will be closer to 1 as long as the regulatory input is low (see **Figure 3C**). On the other hand, if there is no input to the gene, the probability of no change from its previous state is close to 1, and the closeness depends on the parameter d, a small number representing self-degradation. Intuitively, these rules ensure that the state of a gene will flip only if its total input is large enough.

After having calculated these probabilities, the general idea is then to use this information to obtain an appropriate "height" measure for each of the 2^n states. With this in mind, the interest is first in calculating a steady-state probability distribution $P_{SS}(\mathbf{x})$. This stationary probability distribution is analogous to stationary configurations in the deterministic case; however, in the stochastic framework, the probability of being in any particular state, rather than the state of the system, is what is kept invariant along time. In other words, when this stationary distribution is reached, the probability of observing a cell in a particular state does not change. Intuitively, one would expect that attractors would have a higher probability of being reached than transitory states. Thus, from a landscape perspective, the potency of differentiation and height should be inversely related with the probability. The approach that has been followed is to associate this $P_{SS}(\mathbf{x})$ with a height value. Wang has proposed that the probability distribution for a particular state $P(\mathbf{x}_i) = \exp[U(\mathbf{x}_i)]$, and from this expression then $U(\mathbf{x}_i) = -\ln P(\mathbf{x}_i)$, where $i = 1, \ldots, 2^n$. This function U has been termed the (probabilistic) quasi-potential (Huang, 2009, 2012; Wang et al., 2010b)]. How are the "quasi-potential" and the steady-state probability formally related to each other is still an open research area (Zhou et al., 2012) (see below).

The key point which has been emphasized by Wang and coworkers is that, although there is (in general) no potential function directly obtainable from the deterministic equations for a given network, a generalized potential (or "quasi-potential") function can be constructed from its probabilistic description. This generalized potential function is inversely related to the steady-state probability (Wang et al., 2006; Han and Wang, 2007; Lapidus et al., 2008). For the case of the extended BN model, once the transition matrix is calculated, the information of the steady-state probabilities can be obtained by solving a discrete set of master equations (ME) for the network (Han and Wang, 2007). The so-called ME is a dynamical equation for the temporal evolution of a probability distribution (for details, see Haken, 1977; Gardiner, 2009). In discrete form it is written as:

$$\frac{\partial}{\partial t}P(\mathbf{x}_i) = \sum_j W_{ji}P(\mathbf{x}_j) - \sum_j W_{ij}P(\mathbf{x}_i), \tag{12}$$

where we used W_{ij} to denote the transition probabilities resulting from Equation (11). The difference between this dynamical equation and the one discussed in the previous section is that here the time variable is treated as a continuous one. In general, it is quite complicated to analyze MEs. In the case of this model, one ME is obtained for each of the 2^n states. Han and Wang propose to analyzed the whole set of equations following a numerical (iterative) method starting from uniform initial conditions $P_{\mathbf{x}_i}(t_0) = 1/2^n$ and iterating the system until a stationary distribution is reached (Han and Wang, 2007).

1.5. Stochastic EAL Models from Continuous GRNs

As in the case of the deterministic BN model revised above, a general deterministic system of DEs used to describe a GRN can be extended in order to include stochasticity. Such continuous models may be more appropriate to approach certain biological processes. The most intuitive extension considers the introduction of driving stochastic forces. In this approach, Equation (4) is extended to:

$$\frac{d\,x_i(t)}{dt} = F_i(\mathbf{x}, \mathbf{u}) + \xi_i(t), \tag{13}$$

where $\xi_i(t)$ is the ith component of a driving stochastic force with zero mean value (i.e., $< \xi_i(t) >= 0$). This description, the so-called Langevin equation, is frequently used to model cellular dynamics under stochastic fluctuations (Hoffmann et al., 2008; Wang et al., 2010b; Villarreal et al., 2012; Li and Wang, 2013).

Although intuitively simple at first sight, the consideration of a randomly varying quantity affecting the dynamics of the system implies several conceptual issues that should be considered in some detail. Any single cell will follow an erratic trajectory in state space, and its developmental dynamics will make each realization different *even if it starts from exactly the same initial condition*. Under this stochastic scenario, two equivalent perspectives to study the stochastic dynamics can be considered. On the one hand, the analysis could be focused on trajectories described by Langevin-type equations, which describe the developmental dynamics of a single cell (**Figure 4A**). On the other hand, as the stochastic forces $\xi_i(t)$ vary from cell to cell in an *ensemble* (population) of cells, the state $\mathbf{x}(t)$ will also vary from cell to cell at any given time. One therefore may ask for the probability $P(\mathbf{x}, t)$ to find the state of a cell in a given state interval of the state space or, equivalently, for the frequency of cells in the ensemble whose states are in that state interval. In the latter situation, the focus shifts from the dynamics of the state of one cell to the dynamics of the distribution over the states in a given ensemble of cells. Indeed, an equation for the temporal evolution of this distribution $P(\mathbf{x}, t)$ can be constructed, and this corresponds to the so-called *Fokker-Plank equation (FPE)*:

$$\frac{\partial P}{\partial t} = -\sum_i \frac{\partial}{\partial x_i}[A_i(\mathbf{x})P] + \frac{1}{2}\sum_{i,j} Q_{i,j}(\mathbf{x})\frac{\partial^2}{\partial x_i \partial x_j}P. \tag{14}$$

In mathematical terms, the corresponding process is known as a *diffusion process*, a mathematical model for stochastic phenomena evolving in continuous time; the vector $\mathbf{A}(\mathbf{x})$ is known as the drift vector and the matrix $\mathbf{Q}(\mathbf{x})$ as the diffusion matrix (for details, see Risken, 1984; Gardiner, 2009; Fuchs, 2013b). The FPE describes the change of the probability distribution of a cell state during the course of time (**Figure 4B**). Conceptually, the latter modeling perspective can be interpreted as the temporal evolution of a cloud (ensemble) of cells diffusing across the state space following both attracting and stochastic forces (see Huang, 2010 for a conceptual perspective).

The stochastic nature of the trajectories also produce qualitatively richer dynamics in state space. For example, if one is interested in the developmental connection between one specific initial cell state and one specific final cell state—for example, two different given attractors—there is no longer a single

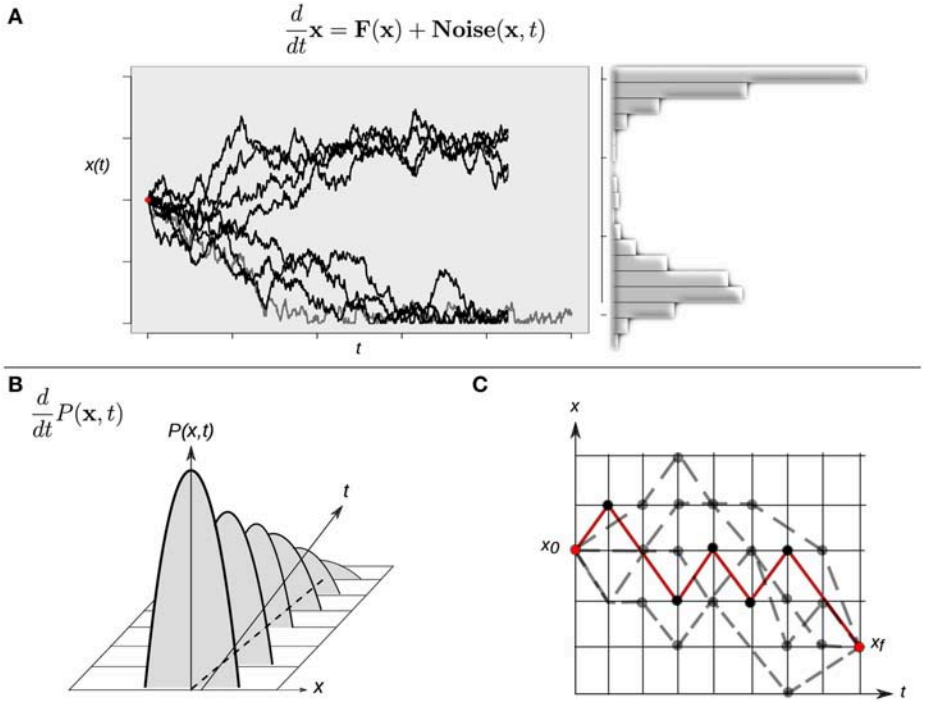


FIGURE 4 | Different approaches to study continuous-time stochastic models of the epigenetic landscape and developmental dynamics. (A) A continuous-time stochastic (diffusion) model is driven by a drift (deterministic) component **F** and an stochastic (**Noise**) force. The graph shows 10 different realizations of the stochastic dynamics of the same, single cell starting form exactly the same initial condition (red dot). This realizations perspective corresponds to the Langevin equation description. The right histogram represents an approximation of the corresponding distribution over the realizations. **(B)** The picture represents the time evolution of a hypothetical probability distribution. A population of cells initially presents a narrow distribution centered at an intermediate state value: most cells have an intermediate state and no individuals show an extreme (low or high) value. As time evolves the shape of the distribution changes—gets wider—, and the population reaches lower and higher values. This perspective corresponds to the Fokker-Planck equation description. **(C)** A cell can follow different paths (gray dotted lines) to reach a final state x_f starting from an initial state x_0. A finer quantitative characterization of the specific transition from state x_0 to state x_f in terms of highly probable paths and difficulty of differentiation processes can be gained by means of calculating a dominant path (red line) for the transition using a path-integral formalism. For simplicity, the cell state is represented by one variable x in all three cases.

possible path connecting them. Instead, the same final cellular phenotype can be reached following different paths in state space (**Figure 4C**). This situation raises yet additional interesting issues: are all the paths equally probable? Is there a dominant path for such a transition from one attractor to another one? Physicists have proposed the so-called path-integral formalism in order to tackle these and similar questions (Wio, 1999). Specifically, one may want to answer what is the probability of starting from an initial cellular phenotype at a certain time and ending in another cellular phenotype at a future time. The conceptual basis of this strategy is based on the idea of calculating an average trajectory (e.g., integrating over the possible paths). The calculated averaged path corresponds to the dominant path that the underlying process is expected to preferentially follow (**Figure 4C**).

Given the intuitive appeal of a landscape perspective to general dynamics, the existence of a potential or "potential-like" function associated with diffusive systems has been an intensive focus of study in theoretical physics and applied mathematics. Ao and co-workers have proposed a transformation that allows the definition of a function $U(\mathbf{x})$ which successfully acquires the dynamical meaning of a potential function. The corresponding approach has been applied successfully to study several biological systems such as the *phage lambda life cycle* (Zhu et al., 2004), and the carcinogenesis processes, Ao et al. (2008), Wang et al. (2013, 2014), and Zhu et al. (2015) from a landscape perspective. This transformation has also been discussed recently in the context of general methods for the decomposition of multivariate continuous mappings $F(\mathbf{x})$ and their associated quasi-potentials (Zhou et al., 2012). From the available decomposition methods, the one that has been applied the most to specific developmental processes is the potential landscape and flux framework proposed by Wang et al. (2008). In this framework, the continuous dynamical mapping $F(\mathbf{x})$ is decomposed into a gradient part and a flux, curl part (for details, see Wang, 2011). This approach has been applied, for example, to the study of the yeast cell cycle [Wang et al. (2006, 2010a)]; a circadian oscillator (Wang et al., 2009); the generic processes of stem cell differentiation and reprogramming (Wang et al., 2010b; Xu et al., 2014); and neural differentiation (Qiu et al., 2012). Recently, this method has been applied in the context of the differentiation and reprogramming of a human stem cell network (Li and Wang, 2013). Here we further discuss the latter as a diffusion landscape approach to study stem cell differentiation.

Although the technical details of decomposition methods for diffusive systems from a landscape perspective are out of the scope of the present review, we point the reader to Ao (2004), Kwon et al. (2005), Yin and Ao (2006), Ao et al. (2007), Ge and Qian (2012), Zhou et al. (2012), and Lv et al. (2014) for further details.

To summarize this section: when a stochastic component with specific properties is introduced in a continuous-time dynamical model of developmental dynamics, the behavior of the system can be studied from different, mathematically equivalent perspectives. One of the perspectives could be more appropriate than the others, given the biological question of interest; the different perspectives complement each other, nonetheless. It is important to note that the three approaches mentioned above (e.g., Langevin, FPE, and path-integral) although just recently introduced in systems biology (Wang et al., 2010b, 2011; Villarreal et al., 2012; Zhang and Wolynes, 2014; Wang et al., 2014); are actually well-established tools in non-equilibrium statistical mechanics and the stochastic approach to complex systems (Haken, 1977; Lindenberg and West, 1990; Gardiner, 2009).

1.6. From EAL Models to Biological Insights

1.6.1. EL Exploration in Flower Morphogenesis

Alvarez-Buylla and collaborators applied the attractor transition probability approach (Equations 5–8 and **Figure 3B**) to explore the EAL explained above in order to study flower patterning shared by most angiosperms or flowering species (Alvarez-Buylla et al., 2008). In flowering plants, a floral meristem is sequentially partitioned into four regions from which the floral organ primordia are formed and eventually give rise to sepals in the outermost whorl, then to petals in the second whorl, stamens in the third, and carpels in the fourth whorl in the central part of the flower. This spatiotemporal pattern is widely conserved among angiosperms. Can the temporal pattern of cell-fate attainment be explained by the interplay of stochastic perturbations and the constraints imposed by a non-linear GRN? Starting from the previously characterized Boolean GRN of organ identity genes in the *A. thaliana* flower (Espinosa-Soto et al., 2004), and applying the stochastic approach described in Equations (5–8), the authors showed that the most probable order in which the attractors are attained is, in fact, consistent with the temporal sequence in which the specification of corresponding cellular phenotypes are observed *in vivo*. The model provided, then, a novel explanation for the emergence and robustness of the ubiquitous temporal pattern of floral organ specification, and also allowed predictions on the population dynamics of cells with different genetic configurations during development (Alvarez-Buylla et al., 2008). Note that in this approach, through the calculation of transition probabilities among attractors, it is possible to explore the EAL associated with a GRN. It also constitutes a new approach to understanding a morphogenic process and also implies that GRN topologies could have, in part, evolved in response to noisy environments. In the same contribution, the authors also showed that a stochastic continuous approximation of the GRN under analysis yielded consistent results. Importantly, in this study it was argued that the fact that observed patterns of cell-fate transitions could be significantly constrained by GRN in the context of noisy perturbations does not excludes the relevance of deterministic signals.

1.6.2. From Probabilistic Landscapes to Putative Cancer Therapies

The probabilistic landscape (quasi-potential) approach has been applied to two specific processes: cell cycle regulation (Han and Wang, 2007), and DNA damage response (Choi et al., 2012). In the former case, the focus was on the global robustness properties of the network. Here we discuss the biological implications derived from the latter case. Choi and collaborators applied this BN probabilistic landscape approach (Equations 9–12 and **Figures 3C,D**) to study state transition in a simplified network of the p53 tumor suppressor protein. The analysis of this network from an EAL perspective allowed the systematic search for combinatorial therapeutic treatments in cancer (Wang, 2013). Given the network, key nodes and interactions that control p53 dynamics and the cellular response to DNA damage were identified by conducting single node and link mutation simulations; as a result, one network component, the molecule Wip1, was identified as one of the critical nodes. The flexibility of the BN model also enabled the specification of a MCF7 cancer cell by fixing the state of three nodes of the "normal" network in the course of simulations (for details, see Choi et al., 2012; Wang, 2013). Having specified two different network models, it was possible to compare the dynamics and associated quasi-potential of both normal and cancer cells in the absence and presence of DNA damage. Previous experimental observations indicated that prolonged p53 activity induces senescence or cell death; this behavior was shown to result from the inhibition of the interaction between the molecules Mdm2 and p53 caused by the action of the small molecule Nutlin-3 (Purvis et al., 2012). Using the model, Choi and collaborators predicted that neither Wip1 nor Mdm2-p53 interaction mutation alone were sufficient to induce cell death for MCF7 cancer cells in the presence of DNA damage; furthermore, the model provided a mechanistic explanation for this behavior: the effect of each of this perturbations alone is not enough to move the system out of an specific attractor's basin. But the simultaneous application of the two perturbations may drive cancer cells to cell death or cell senescence attractors. These theoretical predictions were then validated using single-cell imaging experiments (Choi et al., 2012; Wang, 2013).

This study illustrated in an elegant way how cancer therapeutic strategies can be studied in mechanistic terms using a computational EAL model. It must be pointed out that this result opened the door to the rational design of system dynamics cancer therapeutical techniques, in contrast to trial and error and reductionist approaches that have dominated the biomedical field up to now (Huang and Kauffman, 2013).

1.6.3. A Diffusion Approach to Study the EAL

The three perspectives to study continuous-time stochastic models of developmental dynamics briefly described above and represented in **Figure 4** have been applied to understanding actual developmental cases from an EAL point of view. For example, Villarreal and collaborators recently proposed a procedure to construct a probabilistic EAL by calculating the probability distribution of stable gene expression configurations arising from the topology of a general N-node GRN (Villarreal et al., 2012). In this approach, the focus of study is the temporal evolution

of the distribution over state space (Equation 14 and **Figure 4B**) starting from a position centered on a specific attractor configuration. Intuitively, the proposed framework predicts how a cloud of cells distributed over a particular attractor will diffuse in time to the neighboring regions (attractors) in state space, given a specific GRN (which constraints the state trajectories). The method has been applied to the case of early flower morphogenesis (see subsection above); and its behavior, in both wild type and mutant conditions. The authors recovered patterns that are in agreement with the temporal developmental pattern of floral organs attainment in *A. thaliana* and most flowering species (Alvarez-Buylla et al., 2008; Villarreal et al., 2012). The AEL perspective has recently also given important insights into the problem of carcinogenesis trough the quantitative implementation of the *molecular–cellular network hypothesis* by Ao and co-workers (for details, see Wang et al., 2014; Zhu et al., 2015).

1.6.4. Cell Fate Decisions in the Human Stem Cell Landscape

Recently, Li and Wang adopted the diffusion approach to study a previously published human stem cell developmental network (see Chang et al., 2011) composed of 52 genes (Li and Wang, 2013). In this study they showed how the three perspectives represented in **Figure 4** can complement each other in the study of cellular differentiation: (1) through the numerical analysis of the Langevin-like equations for the complete network they acquired a landscape directly from the statistics of the trajectories of the system (Equation 13 and **Figure 4A**); (2) by means of approximations they studied the evolution of the probabilistic distribution and obtained an steady-state distribution (Equation 14 and **Figure 4B**); and (3) using the path-integral formalism (**Figure 4C**) they calculated the dominant paths (Wang et al., 2011). The obtained paths were interpreted as the biological paths for differentiation and reprogramming (Li and Wang, 2013). As Li and Wang showed, from the results of the three perspectives it is possible to quantitatively describe the underlying EAL. One then may be interested in how the EAL changes in response to specific perturbations.

A general question in stem cell research concerns the underlying mechanisms that explain the known reprogramming strategies, which commonly consist on combining perturbations to specific transcription factors. Li and Wang systematically tested which genes and regulatory interactions imply the greatest alterations to the quantitative properties of the EAL (e.g., height values and transition rates) when perturbed. Interestingly, several biological observations associated with the manipulation of the so-called Yamanaka factors (Oct3/4, Sox2, Klf4, c-Myc)—the transcription factors considered the core regulators in the induction of pluripotency—were consistent with the observed modeling results. For example, simulated knockdown perturbations to these factors consistently increased (lowered) the probability (height) of the differentiation state. On the other hand, the path-integral formalism allowed them to show how specific perturbations to these factors cause the differentiation process to be easier or harder in terms of the time spent during transitions and the characteristics of the differentiation paths. Overall, this study presented an important contribution toward the mechanistic, dynamical explanation of the characterized reprogramming strategies in terms of the properties of the underlying EAL.

1.7. Concluding Remarks

An overall strategy for the practical implementation of what we call EAL models comprises four steps: (1) establishment of an experimentally grounded GRN; (2) characterization of the attractor (and quasi-potential) landscape through dynamical modeling; (3) computational prediction of cell state responses to specific perturbations; and (4) analysis of the prevailing paths of cell fate change. The first step (1) is already a well-established research problem that includes expert curation of experimental data and/or statistical inference. In this review we focused on the second step and presented examples of how steps (3) and (4) can be achieved once a EAL model is effectively constructed. As shown here, there are several ways to implement an EAL model starting from a GRN. The specific choice should be made considering the properties of the network and the associated questions of interest.

The methodologies reviewed here are mostly well-suited to approach the problem of differentiation and temporal cell-fate attainment in a mechanistic setting. The observed behavior results from constraints given by the joint effect of non-linear regulatory interactions and the inherent stochasticity prevalent in GRN. The actual physical implementation of these generic mechanisms in a multicellular system would necessarily imply additional sources of constraint and spatially explicit, multi-level modeling platforms. Tissue-level patterning mechanisms such as cell-cell interactions; chemical signaling; cellular growth, proliferation, and senescence; in addition to mechanic and elastic forces at play in cells, tissues and organs, inevitably impose physical limitations which in turn affect cellular behavior. This would thus imply non-homogenous GRNs with contrasting additional chemical and physical constraints, that in a cooperative manner underlie the emergence of positional information and morphogenetic patterns. Given this fact, the next logical step to extend EAL and associated dynamical models would be to account for these physical processes in an attempt to understand how cellular decisions occur during tissue patterning and not just in cell cultures. Although some progress has been presented in this direction (see, for example Barrio et al., 2010, 2013), the problem remains largely open, specially in terms of explicitly considering the constrains imposed by the underlying GRN and EAL.

From a theoretical perspective, a further challenge would be to carefully evaluate the assumptions implicit in the EAL models. For example, the adoption of the diffusive perspective briefly explained above—which is often taken as a standard in stem cell systems biology—implicitly assumes certain properties about the forces driving the temporal evolution of the system (Lindenberg and West, 1990). Are these conditions universally met by developmental systems? Recent interesting work is starting to suggest the biological relevance of additional constraints such as state-dependent fluctuations (Pujadas and Feinberg, 2012; Weber and Buceta, 2013), as well as time-dependent dynamical behavior (Mitra et al., 2014; Verd et al., 2014). In both cases, a dynamically changing EAL is proposed as a potentially more

accurate description of developmental processes than its static counterpart.

Overall, the application of the methodologies discussed in this review to specific developmental processes has shown the practical relevance of dynamical models consistent with the conceptual basis of the classical EL and the fundamental role of the constraints imposed by the GRN interactions. The different EAL modeling approaches are useful to answer specific questions and can complement each other. So far, EAL models have shown to be an adequate framework for understanding stem cell differentiation and reprogramming events in mechanistic terms; and are also starting to show promise as the basis for rational cancer therapeutic strategies, as well as other interesting issues in developmental biology and evolution.

Acknowledgments

This work was supported by grants from CONACYT, Mexico: 240180, 180380, 167705, 152649 to ERA-B; from UNAM-DGAPA-PAPIIT: IN203113, IN 203214, IN203814 (ERA-B). JD-V receives a Phd scholarship from CONACYT. The authors acknowledge the logistical and administrative help of Diana Romo. The authors acknowledge the Centro de Ciencias de la Complejidad (C3), UNAM.

References

Albert, R., and Othmer, H. G. (2003). The topology of the regulatory interactions predicts the expression pattern of the segment polarity genes in *Drosophila melanogaster*. *J. Theor. Biol.* 223, 1–18. doi: 10.1016/S0022-5193(03)00035-3

Allen, L. J. (2010). *An Introduction to Stochastic Processes with Applications to Biology*. Boca Raton, FL: CRC Press.

Alvarez-Buylla, E. R., Chaos, A., Aldana, M., Benítez, M., Cortes-Poza, Y., Espinosa-Soto, C., et al. (2008). Floral morphogenesis: stochastic explorations of a gene network epigenetic landscape. *PLoS ONE* 3:e3626. doi: 10.1371/journal.pone.0003626

Alvarez-Buylla, E. R., Azpeitia, E., Barrio, R., Benítez, M., and Padilla-Longoria, P. (2010). From abc genes to regulatory networks, epigenetic landscapes and flower morphogenesis: making biological sense of theoretical approaches. *Semin. Cell Dev. Biol.* 21, 108–117. doi: 10.1016/j.semcdb.2009.11.010

Ao, P., Kwon, C., and Qian, H. (2007). On the existence of potential landscape in the evolution of complex systems. *Complexity* 12, 19–27. doi: 10.1002/cplx.20171

Ao, P., Galas, D., Hood, L., and Zhu, X. (2008). Cancer as robust intrinsic state of endogenous molecular-cellular network shaped by evolution. *Med. Hypotheses* 70, 678–684. doi: 10.1016/j.mehy.2007.03.043

Ao, P. (2004). Potential in stochastic differential equations: novel construction. *J. Physics A* 37, L25. doi: 10.1088/0305-4470/37/3/L01

Azpeitia, E., Davila-Velderrain, J., Villarreal, C., and Alvarez-Buylla, E. R. (2014). "Gene regulatory network models for floral organ determination," in *Flower Development: Methods and Protocols*, eds R. José Luis and W. Frank (New York, NY: Springer), 441–469.

Balázsi, G., van Oudenaarden, A., and Collins, J. J. (2011). Cellular decision making and biological noise: from microbes to mammals. *Cell* 144, 910–925. doi: 10.1016/j.cell.2011.01.030

Barrio, R. Á., Hernandez-MacHado, A., Varea, C., Romero-Arias, J. R., and Alvarez-Buylla, E. (2010). Flower development as an interplay between dynamical physical fields and genetic networks. *PLoS ONE* 5:e13523. doi: 10.1371/journal.pone.0013523

Barrio, R. A., Romero-Arias, J. R., Noguez, M. A., Azpeitia, E., Ortiz-Gutiérrez, E., Hernández-Hernández, V., et al. (2013). Cell patterns emerge from coupled chemical and physical fields with Cell proliferation dynamics: the *Arabidopsis thaliana* root as a study system. *PLoS Comput. Biol.* 9:e1003026. doi: 10.1371/journal.pcbi.1003026

Bhattacharya, S., Zhang, Q., and Andersen, M. E. (2011). A deterministic map of Waddington's epigenetic landscape for Cell fate specification. *BMC Syst. Biol.* 5:85. doi: 10.1186/1752-0509-5-85

Chang, R., Shoemaker, R., and Wang, W. (2011). Systematic search for recipes to generate induced pluripotent stem cells. *PLoS Comput. Biol.* 7:e1002300. doi: 10.1371/journal.pcbi.1002300

Choi, M., Shi, J., Jung, S. H., Chen, X., and Cho, K.-H. (2012). Attractor landscape analysis reveals feedback loops in the p53 network that control the cellular response to dna damage. *Sci. Signal.* 5:ra83. doi: 10.1126/scisignal.2003363

Ding, S., and Wang, W. (2011). Recipes and mechanisms of cellular reprogramming: a case study on budding yeast *Saccharomyces cerevisiae*. *BMC Syst. Biol.* 5:50. doi: 10.1186/1752-0509-5-50

Ebeling, W., and Feistel, R. (2011). *Physics of Self-organization and Evolution*. Weinheim: Wiley.com.

Enver, T., Pera, M., Peterson, C., and Andrews, P. W. (2009). Stem cell states, fates, and the rules of attraction. *Cell Stem Cell* 4, 387–397. doi: 10.1016/j.stem.2009.04.011

Espinosa-Soto, C., Padilla-Longoria, P., and Alvarez-Buylla, E. R. (2004). A gene regulatory network model for cell-fate determination during *Arabidopsis thaliana* flower development that is robust and recovers experimental gene expression profiles. *Plant Cell* 16, 2923–2939. doi: 10.1105/tpc.104.021725

Fagan, M. B. (2012). Waddington redux: models and explanation in stem cell and systems biology. *Biol. Philos.* 27, 179–213. doi: 10.1007/s10539-011-9294-y

Ferrell, J. E. (2012). Bistability, bifurcations, and Waddington's epigenetic landscape. *Curr. Biol.* 22, R458–R466. doi: 10.1016/j.cub.2012.03.045

Flöttmann, M., Scharp, T., and Klipp, E. (2012). A stochastic model of epigenetic dynamics in somatic cell reprogramming. *Front. Physiol.* 3:216. doi: 10.3389/fphys.2012.00216

Fuchs, A. (2013a). *Nonlinear Dynamics in Complex Systems*. Berlin; Heidelberg: Springer.

Fuchs, C. (2013b). *Inference for Diffusion Processes: with Applications in Life Sciences*. Berlin; Heidelberg: Springer.

Furusawa, C., and Kaneko, K. (2012). A dynamical-systems view of stem cell biology. *Science* 338, 215–217. doi: 10.1126/science.1224311

Garcia-Ojalvo, J., and Arias, A. M. (2012). Towards a statistical mechanics of cell fate decisions. *Curr. Opin. Genet. Dev.* 22, 619–626. doi: 10.1016/j.gde.2012.10.004

Gardiner, C. W. (2009). *Stochastic Methods*. Berlin: Springer.

Garg, A., Mohanram, K., De Micheli, G., and Xenarios, I. (2012). "Implicit methods for qualitative modeling of gene regulatory networks," in *Gene Regulatory Networks*, eds D. Bart and G. Nele (New York, NY: Springer), 397–443.

Ge, H., and Qian, H. (2012). Landscapes of non-gradient dynamics without detailed balance: stable limit cycles and multiple attractors. *Chaos* 22, 023140–023140. doi: 10.1063/1.4729137

Gilbert, S. F. (1991). Epigenetic landscaping: Waddington's use of cell fate bifurcation diagrams. *Biol. Philos.* 6, 135–154.

González, F., Boué, S., and Belmonte, J. C. I. (2011). Methods for making induced pluripotent stem cells: reprogramming a la carte. *Nat. Rev. Genet.* 12, 231–242. doi: 10.1038/nrg2937

Goodwin, B. C. (1963). *Temporal Organization in Cells. A Dynamic Theory of Cellular Control Processes.* London; New York, NY: Academic Press.

Goodwin, B. C. (2001). *How the Leopard Changed its Spots: the Evolution of Complexity*. Princeton, NJ: Princeton University Press.

Grafi, G. (2004). How cells dedifferentiate: a lesson from plants. *Dev. Biol.* 268, 1–6. doi: 10.1016/j.ydbio.2003.12.027

Graham, T. G., Tabei, S. A., Dinner, A. R., and Rebay, I. (2010). Modeling bistable cell-fate choices in the Drosophila eye: qualitative and quantitative perspectives. *Development* 137, 2265–2278. doi: 10.1242/dev.044826

Haken, H. (1977). *Synergetics. An Introduction. Nonequilibrium Phase Trasitions and Self-organization in Physics, Chemistry, and Biology*. Berlin; Heidelberg; New York, NY: Springer-Verlag.

Han, B., and Wang, J. (2007). Quantifying robustness and dissipation cost of yeast Cell cycle network: the funneled energy landscape perspectives. *Biophys. J.* 92, 3755–3763. doi: 10.1529/biophysj.106.094821

Hoffmann, M., Chang, H. H., Huang, S., Ingber, D. E., Loeffler, M., and Galle, J. (2008). Noise-driven stem cell and progenitor population dynamics. *PLoS ONE* 3:e2922. doi: 10.1371/journal.pone.0002922

Hong, T., Xing, J., Li, L., and Tyson, J. (2012). A simple theoretical framework for understanding heterogeneous differentiation of cd4+ t cells. *BMC Syst. Biol.* 6:66. doi: 10.1186/1752-0509-6-66

Huang, S., and Kauffman, S. (2009). "Complex gene regulatory networks-from structure to biological observables: cell fate determination," in *Encyclopedia of Complexity and Systems Science*, ed R. A. Meyers (New York, NY: Springer), 1180–1293.

Huang, S., and Kauffman, S. (2013). How to escape the cancer attractor: rationale and limitations of multi-target drugs. *Semin. Cancer Biol.* 23, 270–278. doi: 10.1016/j.semcancer.2013.06.003

Huang, S., Guo, Y.-P., May, G., and Enver, T. (2007). Bifurcation dynamics in lineage-commitment in bipotent progenitor cells. *Dev. Biol.* 305, 695–713. doi: 10.1016/j.ydbio.2007.02.036

Huang, S. (2009). Reprogramming cell fates: reconciling rarity with robustness. *Bioessays* 31, 546–560. doi: 10.1002/bies.200800189

Huang, S. (2010). Cell lineage determination in state space: a systems view brings flexibility to dogmatic canonical rules. *PLoS Biol.* 8:e1000380. doi: 10.1371/journal.pbio.1000380

Huang, S. (2011). Systems biology of stem cells: three useful perspectives to help overcome the paradigm of linear pathways. *Philos. Trans. R. Soc. B Biol. Sci.* 366, 2247–2259. doi: 10.1098/rstb.2011.0008

Huang, S. (2012). The molecular and mathematical basis of waddington's epigenetic landscape: a framework for post-darwinian biology? *Bioessays* 34, 149–157. doi: 10.1002/bies.201100031

Huang, S. (2013). Genetic and non-genetic instability in tumor progression: link between the fitness landscape and the epigenetic landscape of cancer cells. *Cancer Metastasis Rev.* 32, 423–448. doi: 10.1007/s10555-013-9435-7

Jaeger, J., and Crombach, A. (2012). "Lifes attractors," in *Evolutionary Systems Biology*, ed O. S. Soyer (New York, NY: Springer-Verlag), 93–119.

Kauffman, S. A. (1969). Metabolic stability and epigenesis in randomly constructed genetic nets. *J. Theor. Biol.* 22, 437–467.

Kauffman, S. (1993). *The Origins of Order: Self Organization and Selection in Evolution*. New York, NY: Oxford University Press.

Kwon, C., Ao, P., and Thouless, D. J. (2005). Structure of stochastic dynamics near fixed points. *Proc. Natl. Acad. Sci. U.S.A.* 102, 13029–13033. doi: 10.1073/pnas.0506347102

Ladewig, J., Koch, P., and Brüstle, O. (2013). Leveling Waddington: the emergence of direct programming and the loss of cell fate hierarchies. *Nat. Rev. Mol. Cell Biol.* 14, 225–236. doi: 10.1038/nrm3543

Lapidus, S., Han, B., and Wang, J. (2008). Intrinsic noise, dissipation cost, and robustness of cellular networks: the underlying energy landscape of mapk signal transduction. *Proc. Natl. Acad. Sci. U.S.A.* 105, 6039–6044. doi: 10.1073/pnas.0708708105

Li, C., and Wang, J. (2013). Quantifying cell fate decisions for differentiation and reprogramming of a human stem cell network: landscape and biological paths. *PLoS Comput. Biol.* 9:e1003165. doi: 10.1371/journal.pcbi.1003165

Li, F., Long, T., Lu, Y., Ouyang, Q., and Tang, C. (2004). The yeast cell-cycle network is robustly designed. *Proc. Natl. Acad. Sci. U.S.A.* 101, 4781–4786. doi: 10.1073/pnas.0305937101

Lindenberg, K., and West, B. J. (1990). *The Nonequilibrium Statistical Mechanics of Open and Closed Systems*. New York, NY: VCH.

Lv, C., Li, X., Li, F., and Li, T. (2014). Constructing the energy landscape for genetic switching system driven by intrinsic noise. *PLoS ONE* 9:e88167. doi: 10.1371/journal.pone.0088167

MacArthur, B. D., Ma'ayan, A., and Lemischka, I. R. (2008). Toward stem cell systems biology: from molecules to networks and landscapes. *Cold Spring Harb. Symp. Quant. Biol.* 73, 211–215. doi: 10.1101/sqb.2008.73.061

MacArthur, B. D., Ma'ayan, A., and Lemischka, I. R. (2009). Systems biology of stem cell fate and cellular reprogramming. *Nat. Rev. Mol. Cell Biol.* 10, 672–681. doi: 10.1038/nrm2766

Mammoto, T., and Ingber, D. E. (2010). Mechanical control of tissue and organ development. *Development* 137, 1407–1420. doi: 10.1242/dev.024166

Mendoza, L., and Alvarez-Buylla, E. R. (1998). Dynamics of the genetic regulatory network for *Arabidopsis thaliana* flower morphogenesis. *J. Theor. Biol.* 193, 307–319.

Mitra, M. K., Taylor, P. R., Hutchison, C. J., McLeish, T., and Chakrabarti, B. (2014). Delayed self-regulation and time-dependent chemical drive leads to novel states in epigenetic landscapes. *J. R. Soc. Interface* 11, 20140706. doi: 10.1098/rsif.2014.0706

Pujadas, E., and Feinberg, A. P. (2012). Regulated noise in the epigenetic landscape of development and disease. *Cell* 148, 1123–1131. doi: 10.1016/j.cell.2012.02.045

Purvis, J. E., Karhohs, K. W., Mock, C., Batchelor, E., Loewer, A., and Lahav, G. (2012). p53 dynamics control cell fate. *Science* 336, 1440–1444. doi: 10.1126/science.1218351

Qiu, X., Ding, S., and Shi, T. (2012). From understanding the development landscape of the canonical fate-switch pair to constructing a dynamic landscape for two-step neural differentiation. *PLoS ONE* 7:e49271. doi: 10.1371/journal.pone.0049271

Risken, H. (1984). *Fokker-Planck Equation*. Berlin; Heidelberg: Springer.

Roeder, I., and Radtke, F. (2009). Stem cell biology meets systems biology. *Development* 136, 3525–3530. doi: 10.1242/dev.040758

Sciammas, R., Li, Y., Warmflash, A., Song, Y., Dinner, A. R., and Singh, H. (2011). An incoherent regulatory network architecture that orchestrates b cell diversification in response to antigen signaling. *Mol. Syst. Biol.* 7:495. doi: 10.1038/msb.2011.25

Sieweke, M. H. (2015). Waddingtons valleys and captain cooks islands. *Cell Stem Cell* 16, 7–8. doi: 10.1016/j.stem.2014.12.009

Slack, J. M. (2002). Conrad hal Waddington: the last renaissance biologist? *Nat. Rev. Genet.* 3, 889–895. doi: 10.1038/nrg933

Strogatz, S. (2001). *Nonlinear Dynamics and Chaos: with Applications to Physics, Biology, Chemistry and Engineering*. New York, NY: Perseus Books Group.

Takahashi, K., and Yamanaka, S. (2006). Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell* 126, 663–676. doi: 10.1016/j.cell.2006.07.024

Takahashi, K., Tanabe, K., Ohnuki, M., Narita, M., Ichisaka, T., Tomoda, K., et al. (2007). Induction of pluripotent stem cells from adult human fibroblasts by defined factors. *Cell* 131, 861–872. doi: 10.1016/j.cell.2007.11.019

Thompson, E. G., and Galitski, T. (2012). Quantifying and analyzing the network basis of genetic complexity. *PLoS Comput. Biol.* 8:e1002583. doi: 10.1371/journal.pcbi.1002583

Verd, B., Crombach, A., and Jaeger, J. (2014). Classification of transient behaviours in a time-dependent toggle switch model. *BMC Syst. Biol.* 8:43. doi: 10.1186/1752-0509-8-43

Villarreal, C., Padilla-Longoria, P., and Alvarez-Buylla, E. R. (2012). General theory of genotype to phenotype mapping: derivation of epigenetic landscapes from N-node complex gene regulatory networks. *Phys. Rev. Lett.* 109:118102. doi: 10.1103/PhysRevLett.109.118102

Von Dassow, G., Meir, E., Munro, E. M., and Odell, G. M. (2000). The segment polarity network is a robust developmental module. *Nature* 406, 188–192. doi: 10.1038/35018085

Waddington, C. H. (1957). *The Strategy of Genes*. London: George Allen & Unwin, Ltd.

Wang, J., Huang, B., Xia, X., and Sun, Z. (2006). Funneled landscape leads to robustness of cell networks: yeast cell cycle. *PLoS Comput. Biol.* 2:e147. doi: 10.1371/journal.pcbi.0020147

Wang, J., Xu, L., and Wang, E. (2008). Potential landscape and flux framework of nonequilibrium networks: robustness, dissipation, and coherence of biochemical oscillations. *Proc. Natl. Acad. Sci. U.S.A.* 105, 12271–12276. doi: 10.1073/pnas.0800579105

Wang, J., Xu, L., and Wang, E. (2009). Robustness and coherence of a three-protein circadian oscillator: landscape and flux perspectives. *Biophys. J.* 97, 3038–3046. doi: 10.1016/j.bpj.2009.09.021

Wang, J., Li, C., and Wang, E. (2010a). Potential and flux landscapes quantify the stability and robustness of budding yeast cell cycle network. *Proc. Natl. Acad. Sci. U.S.A.* 107, 8195–8200. doi: 10.1073/pnas.0910331107

Wang, J., Xu, L., Wang, E., and Huang, S. (2010b). The potential landscape of genetic circuits imposes the arrow of time in stem cell differentiation. *Biophys. J.* 99, 29–39. doi: 10.1016/j.bpj.2010.03.058

Wang, J., Zhang, K., Xu, L., and Wang, E. (2011). Quantifying the Waddington landscape and biological paths for development and differentiation. *Proc. Natl. Acad. Sci. U.S.A.* 108, 8257–8262. doi: 10.1073/pnas.1017017108

Wang, G., Zhu, X., Hood, L., and Ao, P. (2013). From phage lambda to human cancer: endogenous molecular-Cellular network hypothesis. *Quant. Biol.* 1–18. doi: 10.1007/s40484-013-0007-1

Wang, G., Zhu, X., Gu, J., and Ao, P. (2014). Quantitative implementation of the endogenous molecular–cellular network hypothesis in hepatocellular carcinoma. *Interface Focus* 4, 20130064. doi: 10.1098/rsfs.2013.0064

Wang, J. (2011). Potential landscape and flux framework of nonequilibrium biological networks. *Annu. Rep. Comput. Chem.* 7, 1. doi: 10.1016/B978-0-444-53835-2.00001-8

Wang, W. (2013). Therapeutic hints from analyzing the attractor landscape of the p53 regulatory circuit. *Sci. Signal.* 6, pe5. doi: 10.1126/scisignal.2003820

Weber, M., and Buceta, J. (2013). Stochastic stabilization of phenotypic states: the genetic bistable switch as a case study. *PLoS ONE* 8:e73487. doi: 10.1371/journal.pone.0073487

West-Eberhard, M. J. (2003). *Developmental Plasticity and Evolution*. New York, NY: Oxford University Press.

Wio, H. (1999). "Application of path integration to stochastic processes: an introduction," in *Fundamentals and Applications of Complex Systems*, ed Nueva (San Luis: Univ. UN), 253.

Xu, L., Zhang, K., and Wang, J. (2014). Exploring the mechanisms of differentiation, dedifferentiation, reprogramming and transdifferentiation. *PLoS ONE* 9:e105216. doi: 10.1371/journal.pone.0105216

Yin, L., and Ao, P. (2006). Existence and construction of dynamical potential in nonequilibrium processes without detailed balance. *J. Phys. A* 39, 8593. doi: 10.1088/0305-4470/39/27/003

Zhang, B., and Wolynes, P. G. (2014). Stem cell differentiation as a many-body problem. *Proc. Natl. Acad. Sci. U.S.A.* 111, 10185–10190. doi: 10.1073/pnas.1408561111

Zhou, J. X., and Huang, S. (2011). Understanding gene circuits at cell-fate branch points for rational cell reprogramming. *Trends Genet.* 27, 55–62. doi: 10.1016/j.tig.2010.11.002

Zhou, J. X., Aliyu, M., Aurell, E., and Huang, S. (2012). Quasi-potential landscape in complex multi-stable systems. *J. R. Soc. Interface* 9, 3539–3553. doi: 10.1098/rsif.2012.0434

Zhu, X.-M., Yin, L., Hood, L., and Ao, P. (2004). Calculating biological behaviors of epigenetic states in the phage λ life cycle. *Funct. Integr. Genomics* 4, 188–195. doi: 10.1007/s10142-003-0095-5

Zhu, X., Yuan, R., Hood, L., and Ao, P. (2015). Endogenous molecular-cellular hierarchical modeling of prostate carcinogenesis uncovers robust structure. *Prog. Biophys. Mol. Biol.* 117, 30–42. doi: 10.1016/j.pbiomolbio.2015.01.004

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Chapter 4

Resultados

No research program has sought to determine the implications of adaptive processes that mold systems with their own inherent order.
— Stuart Kauffman, *The Origins of Order* (1993)

... all organisms are a mixture of conserved and nonconserved processes (said otherwise, of unchanging and changing processes), rather than a uniform collection of processes that change equally in the sources of variation in the course of evolution.
— Kirschner and Gerhart, *The Plausibility of Life* (2005)

Davila-Velderrain *et al. BMC Systems Biology* (2015) 9:20
DOI 10.1186/s12918-015-0166-y

BMC Systems Biology

RESEARCH ARTICLE Open Access

Reshaping the epigenetic landscape during early flower development: induction of attractor transitions by relative differences in gene decay rates

Jose Davila-Velderrain[1,2], Carlos Villarreal[2,3*] and Elena R Alvarez-Buylla[1,2*]

Abstract

Background: Gene regulatory network (GRN) dynamical models are standard systems biology tools for the mechanistic understanding of developmental processes and are enabling the formalization of the epigenetic landscape (EL) model.

Methods: In this work we propose a modeling framework which integrates standard mathematical analyses to extend the simple GRN Boolean model in order to address questions regarding the impact of gene specific perturbations in cell-fate decisions during development.

Results: We systematically tested the propensity of individual genes to produce qualitative changes to the EL induced by modification of gene characteristic decay rates reflecting the temporal dynamics of differentiation stimuli. By applying this approach to the flower specification GRN (FOS-GRN) we uncovered differences in the functional (dynamical) role of their genes. The observed dynamical behavior correlates with biological observables. We found a relationship between the propensity of undergoing attractor transitions between attraction basins in the EL and the direction of differentiation during early flower development - being less likely to induce up-stream attractor transitions as the course of development progresses. Our model also uncovered a potential mechanism at play during the transition from EL basins defining inflorescence meristem to those associated to flower organs meristem. Additionally, our analysis provided a mechanistic interpretation of the homeotic property of the ABC genes, being more likely to produce both an induced inter-attractor transition and to specify a novel attractor. Finally, we found that there is a close relationship between a gene's topological features and its propensity to produce attractor transitions.

Conclusions: The study of how the state-space associated with a dynamical model of a GRN can be restructured by modulation of genes' characteristic expression times is an important aid for understanding underlying mechanisms occurring during development. Our contribution offers a simple framework to approach such problem, as exemplified here by the case of flower development. Different GRN models and the effect of diverse inductive signals can be explored within the same framework. We speculate that the dynamical role of specific genes within a GRN, as uncovered here, might give information about which genes are more likely to link a module to other regulatory circuits and signaling transduction pathways.

Keywords: Gene regulatory network, Epigenetic landscape, Attractor landscape, Differentiation, Flower development, Attractor transitions

*Correspondence: carlos@fisica.unam.mx; eabuylla@gmail.com
[2] Centro de Ciencias de la Complejidad (C3), Universidad Nacional Autónoma de México, Cd. Universitaria, 04510 México, D.F., México
[1] Instituto de Ecología, Universidad Nacional Autónoma de México, Cd. Universitaria, 04510 México, D.F., México
Full list of author information is available at the end of the article

Background

The *systems* perspective to biology has successfully rephrased long-standing questions in developmental biology in terms of the dynamical behavior of molecular networks [1-4]. A salient example is the increasing use of gene regulatory network (GRN) models to study cell-fate specification [5-9]. How can cells with the same genotype and gene regulatory network in multicellular organisms attain different cell fates? How are the steady-state gene expression configurations that characterize each cell-type attained? Why do we observe certain cellular phenotypes and not others? How are the temporal and spatial patterns of cell-fate decisions established and how are they robustly maintained? The dynamical analysis of GRNs has given insights into these and other important questions concerning cell differentiation and morphogenesis, the two components of development. In short, GRN models are showing how observed differentiation patterns can be understood in mechanistic terms [10]. Overall, experimentally grounded GRN models constitute multistable dynamical systems able to recover stable steady states (or *attractors*) corresponding to fixed profiles of gene activation that mimic those characterizing different cell types in both plants and animals (e.g., [11,12]). Such profiles are commonly interpreted as cell fates [1,4,13].

The first, and arguably the simplest, model of GRN dynamics is the Boolean network model proposed by Stuart Kauffman [14]. This model is based on strong assumptions, mainly: (1) gene activity shows binary (on/off) behavior; (2) the temporal change in gene activity occurs in discrete, regular steps; and, originally, (3) the activity state of the whole network evolves in a synchronized manner [15]. Albeit highly abstract at first sight, the applicability of Boolean GRNs, as well as derived conceptual implications, have been supported extensively both by experimental observations [5,16,17] and by theoretical GRNs grounded on experimental data [11,18]. A first example of the latter was proposed to understand cell-fate attainment during early flower development [19]. The Boolean GRN model has become a well established modeling tool in systems biology that is intuitive and attractive to biologists [20,21].

In addition, simple GRN dynamical models are enabling the formalization of old biology metaphors such as the conceptual model of the epigenetic landscape (EL) proposed by C.H. Waddington in 1950s [22-25]. In modern post-genomic biology the EL has been consolidated as the preferred conceptual framework for the discussion of the mechanistic basis underlying cellular differentiation and plasticity [26-28]. A formal basis for this metaphorical EL is being developed in the context of GRNs [24,29-32]. The key for this formalization is to consider that, as well as generating the cellular phenotypic sates (attractors), the GRN dynamics also partitions the whole state-space – the abstract space containing all the possible states of a given system – in specific regions restricting the trajectories from one state to another one. The formalization of the EL in this context is conceptually straightforward: the number, depth, width, and relative position of the attractor's basins of attraction would correspond to the hills and valleys of the metaphorical EL [24]. Here, we refer to the structured order of the basins in state-space as the attractors landscape (AL). For our purposes, the characterization of an AL would correspond, in practical terms, to the characterization of an EL (see below). There is an increasing interest to model the EL associated with a GRN [9,24,30,33-37].

Despite developments in both the conceptual and technical aspects of GRN modeling, interest in novel questions associated with developmental cell plasticity calls for extended modeling frameworks. For example, previous modeling approaches are not able to address the importance of quantitative alterations of the GRN components in attractors (cell-fates) attainment and transitions, or the importance of particular GRN components in moving the system from a particular steady-state or cell fate to another one. In an attempt to contribute to such a need, in this work we propose a modeling framework that integrates standard dynamical systems analyses to extend the simple GRN Boolean model in order to address questions regarding the impact of gene specific perturbations in cell-fate decisions during development. Two different, non-exclusive, approaches are commonly followed in the study of GRN developmental dynamics: (1) analyzing a large set of randomly (or exhaustively) assembled networks (see, for example [38-40]); or (2) focusing on one, well-characterized and experimentally grounded GRN [11,18]. In this work we adopt the second approach.

One of the first GRN models, which is experimentally grounded and has been extensively validated and used to test different approaches, is the floral organ specification GRN (FOS-GRN). The GRN model proposes a regulatory module underlying floral organ determination in *Arabidopsis thaliana* during early stages of flower development [11,19,41]. The network is grounded in experimental data for 15 genes and their interactions. Among the 15 genes, five are grouped into three classes (A-type, B-type, and C- type), whose combinations have been shown - through molecular developmental genetic studies - to be necessary for floral organ cell specification. A-type genes (AP1 and AP2) are required for sepal identity, A-type together with B-type (AP3 and PI) for petal identity, B-type and C-type (AGAMOUS) for stamen identity, and the C-type gene (AG) alone for carpel primordia cell identity. The so-called ABC model describes such combinatorial activities during floral organ determination [42]. The original Boolean FOS-GRN converges to ten attractors that correspond to the main cell types observed

during early flower development, and thus provided a mechanistic explanation to the ABC model. Six attractors correspond to sepal (Sep), petal (Pt1 and Pt2), stamen (St1 and St2), and carpel (Car) primordial cells within flower meristems with the expected ABC gene combinations for each floral organ primordi. In addition it explained the configurations that characterize the inflorescence meristem: four attractors correspond to meristematic cells of the inflorescence, which is partitioned into four regions (Inf1, Inf2, Inf3, and Inf4). This network has become one of the prototypical systems for theoretical analyses of cell differentiation and morphogenesis [43], and it has been shown to be well-suited to explore new questions and propose new methodologies.

For example, recently an EL model for flower development based on a continues stochastic approximation of the Boolean GRN showed that characteristic multigene configurations emerge from the constraints imposed by the GRN; but the temporal pattern of cell transitions also seems to depend on the asymmetry in gene expression times-scales for some of the main regulators [33]. Based on this work, it was suggested that parameters representing finer regulatory processes, such as gene expression decay rates, enable richer and more accurate descriptions of the underlying cellular transitions. Specifically, the results suggested that relative differences in the decay rates of particular genes may be important for the establishment of the robust pattern of differentiation transition observed during floral organ determination. Thus, along with the constraints imposed by the GRN, a hierarchy of decay times of gene expression may define alternative routes to cell fates [21,33]. This possibility has not been studied systematically yet and it might prove crucial to undertand how such GRN modules are connected to signal transduction pathways that alter cell-fate attainment patterns.

Given the background exposed above, *a first question concerns the systematic exploration of the effect of a hierarchy of gene expression times on cell-fate specification during early flower development*. On the other hand, flower developmental mechanisms have been shown to result largely from the global self-organizational properties of the FOS-GRN; yet, it has not been straightforward to establish differences in the functional (dynamical) role of individual genes within the network. Therefore, *a second question concerns whether by analyzing gene dynamics we can test if there are such differences and, if so, if they correlate with biological observables*. Given that both questions require modeling exercises that go beyond a simple Boolean GRN model, in this contribution we first propose a modeling framework to extend the Boolean FOS-GRN model to a continuous system, and then show how it can be used to explore the questions addressed here.

For the sake of concreteness, we frame the questions in the context of the dynamics of early flower development as follows: (1) We define the propensity of the Boolean stationary gene configuration to be transformed by changes of particular gene parameters as a proxy for gene functional role. (2) We test as a control parameter the genes characteristic decay rate in order to further explore the hypothesis raised in [33], that differences in gene decay rates may potentially guide cell-fate decisions during flower development. (3) We contrast the dynamical/biological classification with the known experimental data regarding the role of the ABC genes. In other words, we functionally classify the genes in the network by exploring their propensity to produce qualitative changes in the AL that would ultimately lead to cell-fate decisions (*i.e.,* attractor transitions). We also analyze the robustness of each attractor by means of their propensity (or lack thereof) to undergo such induced transitions. We hypothesize that there is a relationship between the impact of specific genes in the dynamics of the whole GRN, their biological function, and the observed hierarchy of differentiation events during early flower development.

Overall, this work constitutes a first step towards the dynamical, mechanistic characterization of the main molecular regulators of flower development; and provides a general methodological framework to approach similar questions in other developmental processes. It also provides hypotheses concerning which genes within the FOS-GRN are more likely to link this module to other regulatory circuits and signaling transduction pathways which might be crucial for the temporal progression of flower development. In conclusion, the approach put forward here allows analyses of the role of the genes' decay rates in modifying the AL and thus affecting cell-fate transitions or patterning.

Methods

Modeling framework

The scope of biological questions that Boolean GRN models are suited to address can be expanded. Here we focus on two specific questions that are important for developmental biology and which cannot be addressed by Boolean models – as originally proposed. (1) Although gene knockout or over-expression experiments are straightforward to simulate using a Boolean model, the richness of gene interactions may be more thoroughly explored by considering the intertwined dynamics of differentiation stimuli (microambient alterations, chemical signaling, catalytic reactions, etc.) and gene characteristic expression times which determine the developmental process itself, and which are not easily taken into account in a Boolean approach due to the absence of genes' specific parameters. (2) It is not straightforward to study potential transition events among the already characterized stable

cellular phenotypes with the Boolean deterministic formalism. With this limitations in mind, here we propose a novel modeling framework as an extension of the original Boolean GRN model. Our goal was to devise an extended methodology able to circumvent these limitations while maintaining the simplicity and clarity of the Boolean model. The proposed framework includes the following steps (see Figure 1): (1) the characterization of the dynamical behavior of an experimentally grounded Boolean GRN - and its associated AL, (2) the transformation of the Boolean model into a system of ordinary differential equations (ODEs) with an equivalent AL, (3) an attractor-wise, gene-wise numerical bifurcation analysis using the characteristic decay rate of each gene as a control parameter [43,44], and (4) the classification of genes into groups according to their propensity to induce qualitative changes to the AL and their potential to cause specific transitions between attractors.

Boolean GRN model

A Boolean network is a dynamical model with discrete time and discrete state variables. This can be expressed formally as:

$$x_i(t+1) = F_i(x_1(t), x_2(t), \ldots, x_k(t)), \tag{1}$$

where the set of functions F_i are logical prepositions (or truth tables) expressing the relationship between a gene i and its k regulators, and where the state variables $x_i(t)$ can take the discrete values 1 or 0 indicating whether the gene i is expressed or not at a certain time t, respectively. An experimentally grounded Boolean GRN model is completely specified by the set of genes proposed to be involved in the process of interest and the associated set of logical functions derived from experimental data [21]. The set of logical functions for the FOS-GRN used in this study is included in Additional file 1. The dynamical analysis of the Boolean network model was conducted using the package *BoolNet* [45] within the *R* statistical programming environment (www.R-project.org).

Continuous GRN model

In order to characterize qualitative changes in the dynamics of the GRN under continuous variations of a given parameter (here a gene's decay rate) we study a continuous representation of the discrete Boolean dynamics. Several approaches have been used to describe a Boolean GRN as a continuous system [21,33,46,47]. Here we adopt a system of ODEs of the form:

$$\frac{dx_i}{dt} = \Theta[f_i(x_1, x_2, \ldots, x_k)] - k_i x_i, \tag{2}$$

where k_i represents the expression decay rate of the gene i of the GRN. The function f_i results from performing a transformation to the corresponding boolean function F_i following the rules:

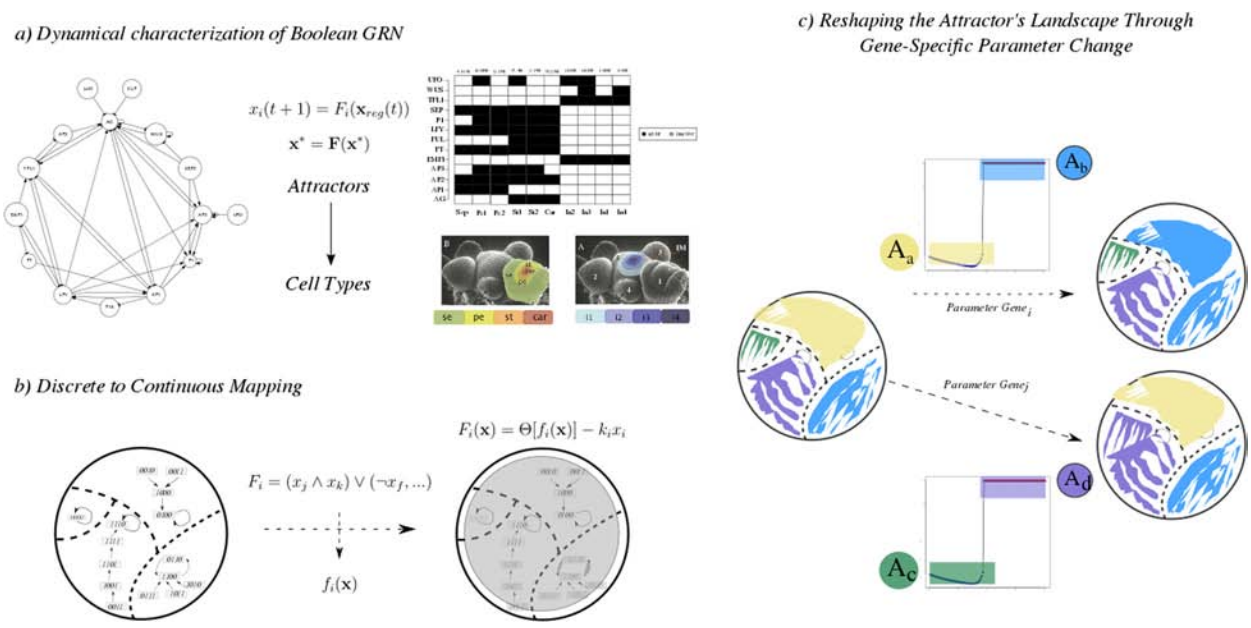


Figure 1 Schematic representations of the modeling methodology. **a)** The starting point is an experimentally grounded and dynamically characterized GRN Boolean model. Here the FOS-GRN is used, which recovers ten fixed-point attractors representing the cell-types observed during early flower development. **b)** The Boolean model is transformed into an equivalent continuous dynamical model. A set of rules is applied to the logical propositions of the Boolean model in order to derive a logic-based ODE model in continuous state-space. **c)** An attractor-wise, gene-wise numerical bifurcation analysis is performed. Because of qualitative changes to the AL induced by increasing parameter values several basins of attraction may merge into one, causing an inevitable cell-fate decision (*i.e.,* an attractor transition).

$$\begin{aligned} x_i(t) \wedge x_j(t) &\rightarrow x_i(t) \,.\, x_j(t), \\ x_i(t) \vee x_j(t) &\rightarrow x_i(t) + x_j(t) - x_i(t) \,.\, x_j(t), \\ \neg x_i(t) &\rightarrow 1 - x_i(t). \end{aligned} \quad (3)$$

Following [21,33] we consider that the input-response function associated to each gene displays a saturation behavior characterized by a logistic function. In this case, the input associated with the gene i takes the form:

$$\Theta[f_i(x_1, x_2, \ldots, x_k)] = \frac{1}{1 + \exp[\,-b[f_i(x_1, x_2, \ldots, x_k) - \epsilon]\,]}, \quad (4)$$

where ϵ is a threshold level (usually $\epsilon = 1/2$), and b the input saturation rate. For $b >> 1$, the input function displays dichotomic behavior. A stationary state is defined by $dx_i/dt = 0$, so that Eq.(2) yields

$$x_i^s = \frac{1}{k_i} \Theta \left[f_i \left(x_1^s, x_2^s, \ldots, x_k^s \right) \right], \quad (5)$$

where x_i^s denotes the stationary value. We observe that the expression level of the GRN node i is inversely proportional to its decay rate, so that for a fast decay rate $k_i \gg 1$ the expression level $x_i^s \rightarrow 0$, while for a slow decay $k_i \ll 1$, $x_i^s \gg 1$. Thus, a hierarchy in gene decay rates determines a pattern of relative gene expression levels.

The obtained system of ODEs is included on Additional file 1. Similar logic-based ODE models have been presented before (see, for example [48,49]). The numerical analysis of the system of ODEs was conducted using inhouse *R* code exploiting the functions provided in the packages *deSolve* [50] and *rootSolve* [50], as described in [51]. During preliminary simulation experiments we observed that under the specified parameter values the uncovered fixed-point attractors always showed extreme values – *i.e.,* close to either 0 or 1, but not to 0.5.

Attractors landscape operational definition

The Attractors Landscape (AL) is specified by the exhaustive characterization of the state-space. We operationally define the AL as the data structure containing two elements: (1) a $2^n \times n$ state-space matrix, a matrix whose rows correspond to each of the 2^n possible states of a Boolean GRN; and (2) a vector of length 2^n whose elements take values A_i from the set $\{1, \ldots, A_n\}$ where A_n is the number of attractors of a given Boolean Network. This structure thus maps each state to its corresponding attractor. For the case of the ODEs model, the obtained attractor states were discretized in order to have a direct comparison with the Boolean model. Following [52] an unsupervised k-means clustering algorithm [53] with two clusters (*i.e.,* $k = 2$) corresponding to the two binary values was used for the discretization task (for details see [52]).

Bifurcation analysis

All bifurcation analyses were conducted numerically using the following algorithm. A specific attractor is taken as an initial condition in an ODEs initial-value problem. For each active gene in the attractor state: (1) an ordered set of values for the control parameter (here the gene's decay rate k_i) is chosen – while the rest of the parameters are kept constant; (2) the ODEs are solved numerically until reaching an steady state, each time using a different parameter value, and for all the parameter values in the set; and (3) a plot is generated with parameter values in the x-axis and the total sum y of the single gene expression values for the n genes (*i.e.,* $y = \sum_{i=1}^{n} x_i^*$) of the obtained steady state x_i^* in the y-axis. The analysis is performed for each attractor. Qualitative changes are identified by the occurrence of sudden jumps in the bifurcation graphs.

Data analysis

Network topology

For each gene (node) in the FOS-GRN the following measures of topological importance were calculated: degree (number of nodes it is connected to), in-degree (number of connections directed towards it), out-degree (number of connections directed towards other nodes), and betweenness (fraction of all shortest paths that pass through it). All network topological computations were conducted using the *igraph* package [54]. In order to test for the association of the genes propensity to produce AL qualitative changes and their topological features within the network, simple linear regression models were fitted using the calculated propensity of each gene to produce a qualitative change as response variable and each topological feature as predictor.

To test whether interacting genes in the FOS-GRN have a related propensity to produce AL alterations in response to an increase in their decay rate. The average absolute difference of the value of the calculated gene sensitivity between interacting components in the network was calculated and then used as a statistic in a simulation (sampling) procedure in order to assess how frequently it is expected to observe this or a smaller value in an ensemble of similar but random networks. Specifically, 100,000 networks each with the same number of nodes and interactions were generated, and the statistic was calculated for each of these networks. The estimated distribution of the statistic over the ensemble of networks was then used to calculate the probability of observing a value equal or smaller than that calculated in the FOS-GRN.

Results

Dynamical analysis of the GRN

The GRN underlying early flower development (refered to as FOS-GRN) was used as a study case. The most recent version reported in [33] was used. The corresponding

logical update rules are reported in Additional file 1. The first task was to characterize the GRN dynamical behavior and its associated AL. The global dynamical behavior of the network was analyzed by the exhaustive characterization of all steady states using all possible initial conditions. Specifically, we calculated its attractor states and their corresponding basins of attraction. We arranged both initial conditions and corresponding attractor into an AL structure (see methods). As expected, the network recovered 10 fixed-point attractors: four corresponding to the four regions of the inflorescence meristem (Inf1, Inf2, Inf3, and Inf4), and six to the four floral organ primordial cells within the flower primordia (Sep, Pt1, Pt2, St1, St2, and Car). The two attractors corresponding to petals (Pt1 and Pt2) are identical except for the state of activation of the UFO gene, and the same holds for the two stamen attractors (St1 and St2). The attractors and its basins are reported in Additional file 1. We then transformed the Boolean network into a system of ODEs (see Methods).

A series of studies have extensively validated the Boolean FOS-GRN model in terms of increasingly available experimental data; for example, it has been shown that its dynamical behavior is robust enough as to predict the experimentally induced phenotypes in several mutant conditions [11,19,24,55]. In order to preserve such validated behavior we derived a ODEs model preserving the attractors and basins of attraction uncovered in the Boolean case. The input-response function included in the proposed continuous model contains 2 parameters: b, and ϵ. The value of the parameter b was chosen as the smallest integer value able to recover the same fixed-point attractors and their basins of the Boolean model. We tested a range of values $b = i$ for $[1, ..i.., 40]$. We found that a value of $b \geq 5$ is able to recover the same attractors and basin sizes that the ones uncovered with the Boolean model. We use a value of $b = 5$ for all the following calculations. The ϵ parameter is a threshold level, for simplicity a value of $\epsilon = 0.5$ was used. For this first analysis the decay parameter for each gene was set to $k_i = 1$. The 10 attractors obtained with these settings, and its basins size are shown in Additional file 1. Thus, we derived two dynamical models for the FOS-GRN with an equivalent behavior in terms of the uncovered attractors and basins of attraction. We specified an AL structure for each model.

Bifurcation analysis

We performed a numerical analysis in order to explore the propensity of single genes to qualitatively change the attractor states where they are expressed (and thus induce attractor transitions in the AL) in response to an increase in their decay rate parameter (see Methods). To illustrate our analyses, we generated a set of graphs, one per each gene expressed in each attractor. In the graph we plotted the initial attractor state and its progressive change resulting from altering the decay parameter k_i. If m genes were active in the attractor in question, the analysis was conducted for each gene i for $i = [1, \ldots, m]$. We performed the analysis to each attractor j for $j = [1, \ldots, 10]$. Figure 2 shows the graphs obtained for the genes corresponding to carpels (Car) attractor. In this case, only the genes AG and LFY were able to induce an phase transition. Whereas gene AG produces a transition between already characterized attractor states ($Car \rightarrow Sep$), the change in LFY produces a new attractor state. The graphs for all the attractors (and their genes) are reported in Additional file 2. We found that for each attractor at least one of its expressed genes is able to produce a qualitative change to the AL. Some genes (attractors) are more likely to produce (undergo) attractor transitions. These results suggest that, by systematically testing the potential of altering specific genes qualitatively changes the GRN underlying AL, we can uncover differences in the genes functional (dynamical) role in the overall system under analysis.

Gene classes

In order to have a better understanding of the nature of the uncovered differential functional (dynamical) role of genes, we classified the genes according to their propensity to induce attractor transitions. Table A1, in Additional file 1 summarizes the result of all the bifurcation analyses. For each attractor, and for each perturbed gene, we registered whether a qualitative change is produced or not, and the final attractor attained after the simulated change. In order to numerically express the propensity of each gene to induce qualitative changes, we counted the number of times a gene is able to produce a qualitative change and normalized this number by the number of times the gene is expressed among the 10 attractors. The resulting scale is shown in Figure 3. We will refer to this quantified propensity to induce qualitative changes (*phase transitions*) as the metric *PT*. In order to classify a gene with either high or low propensity, we clustered the genes described by the quantified propensity *PT* in two groups using the k-means clustering algorithm [56]. According to this analysis, the genes with higher propensity are: UFO, AP1, WUS, AG, TFL1, EMF1, and LFY (see Figure 3). On the other hand, genes were also classified depending on whether or not, when they induce a qualitative change, are able to induce a transition between already characterized attractor states. The genes found to be able to produce this type of transitions are: UFO, AP1, WUS, AP3, AG, TFL1, EMF1, and PI. Additionally, we also classified the genes depending on whether or not they are able to produce new attractor states after the qualitative change. The genes that show this behavior are: SEP, AP2, PI, LFY. The three classes are shown in Table 1. In Figure 4 we map to each node in the graph of the GRN its corresponding metric *PT*.

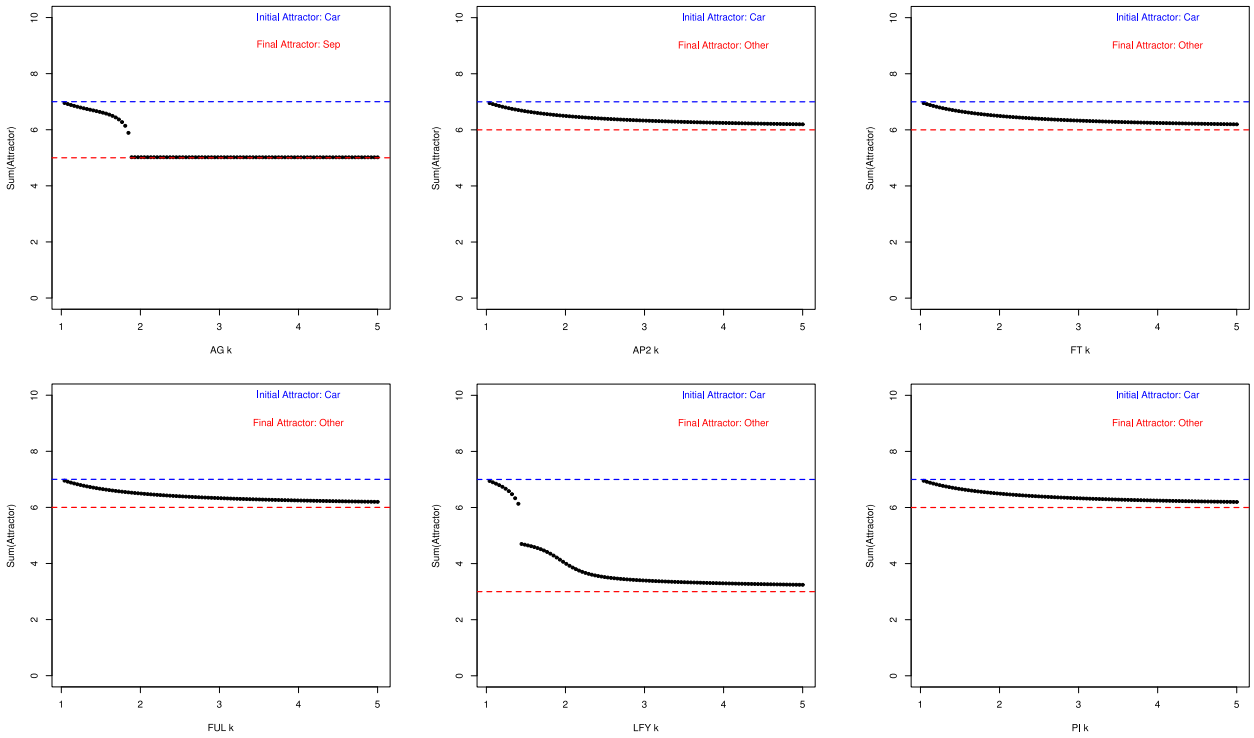


Figure 2 Bifurcation diagrams. Graphs obtained as a result of the Bifurcation analysis performed for the genes corresponding to carpels (Car) attractor. The genes AG and LFY were able to induce qualitative changes. The gene AG produced a transition between already characterized attractor states (Car → Sep). The change in LFY produced a new attractor state.

Analysis of the classes of genes

In order to test if there is evidence of an association between the differential functional role of genes and background biological knowledge, we compared the representation of the ABC genes and the additional (non-ABC) genes of the FOS-GRN within each of the classes described in the previous subsection, and listed these in Table 1. We followed two procedures: (1) calculated the gene frequency of each biological group (e.g. ABC, or Additional) within each gene class, (2) perform a hypergeometric test for biological group over-representation. Figure 5a shows the results. We found the following patterns. In the classes defined by the gene propensity to induce qualitative changes, there is a lower (higher) representation of ABC genes in the high (low) propensity class with respect to the other additional genes. On the other hand, in the classes defined by the gene capacity to produce attractor transitions between known or unknown attractors, there is a higher representation of ABC genes with respect to the other additional genes in both classes. These results suggest that ABC genes are less likely to produce qualitative changes in the AL by induced changes in their expression dynamics - at least under a relatively higher decay rate as tested here - than the non-ABC genes in the network. On the other hand, if such a qualitative change occurs, ABC genes are more likely to both induce inter-attractor transitions and to specify novel attractors than the non-ABC genes in the network. These seemingly contradictory results can be understood by taking into consideration the relative robustness of the different attractors against such parameter perturbations (see below).

Attractors propensity to undergo transitions

Taking in consideration that not all the genes are expressed in all the attractors, we also compared the propensity of the different attractors to undergo attractor transitions by calculating the frequency of attractor transitions per attractor as the number of undergone attractor transitions normalized by the number of genes expressed in the respective attractor. The results are shown in Figure 5b. For this analysis we mapped all the states in the AL corresponding to any of the four inflorescence attractors (Inf1, Inf2, Inf3, and Inf4) into a single *Inf* attractor. We also mapped the states of the attractors (St1, St2) and (Pt1, Pt2) to the individual attractors *St* and *Pt*, respectively. Hence, the system had a total of five attractors. We found that the inflorescence attractor is the attractor with the highest propensity. Specifically, a relatively higher decay rate of any of the genes expressed in the inflorescence attractors (TFL1, EMF1, UFO, WUS) with respect to the other genes always produces an attractor transition.

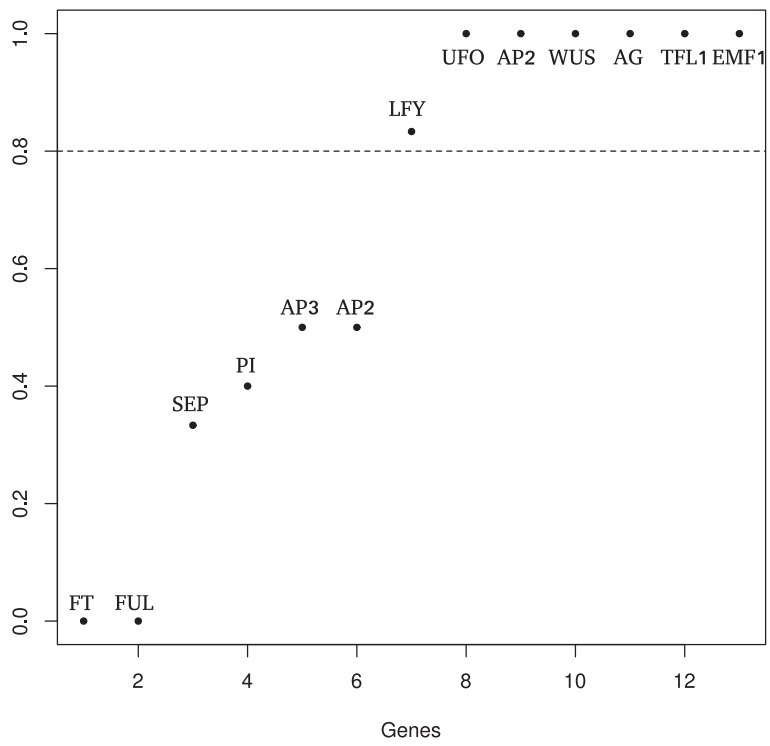


Figure 3 *PT* metric values. The plot shows propensity to induce phase transitions quantified for each gene. The horizontal line divides the genes into groups of higher (above) of lower (below) propensity. The two groups are based on a clustering analysis.

Three of the flower attractors (Car, Sep, St) show a frequency of attractor transitions lower than 0.5, while the remaining flower attractor (Pe) shows a frequency ~ 0.6. These results suggest a relationship between the propensity of undergoing attractor transitions and the direction of differentiation during early flower development - being less likely to induce attractor transitions as the course of development progresses, or to produce a reprogramming from a floral organ attractor to an inflorescence one. Interestingly, attractors propensity to undergo attractor transitions do not correlate with the attractors basin sizes (see Figure 5b), as intuitively expected.

Table 1 Gene classes according to their propensity to produce qualitative changes to the attractors

Classification	Genes
High propensity genes	UFO, AP1, WUS, AG, TFL1, EMF1, LFY
Low propensity genes	SEP, FT, AP3, AP2, PI, FUL
Genes causing transition between known attractors	UFO, AP1, WUS, AP3, AG, TFL1, EMF1, PI
Genes causing transition between unknown attractors	SEP, AP2, PI, LFY

Genes propensity to produce qualitative changes and network structure

Given that it is common to provide evidence of the gene importance in the context of networks by considering only each gene's topological features [57], we tested if the gene's propensity to produce qualitative changes to the AL as defined here is correlated with topological properties. Specifically, we tested an association between each of genes topological features and the quantified gene's propensity of producing a qualitative change to the AL (*PT* metric) by performing linear regression analyses. We characterized each node by a set of network topological features, which express numerically the placement of each gene within the network. For each gene (node) in the FOS-GRN we calculated two commonly used measures of topological importance: degree (number of nodes it is connected to), and betweenness (fraction of all shortest paths that pass through it). We also considered that the dynamical behavior of the GRN is associated with the type of interactions within the network, thus we specified further the degree feature into in-degree or out-degree. Interestingly, we found a significant relationship between *PT* metric and two predictor variables: out-degree and betweenness (p-value $= 0.03$). In Figure 4 we represent

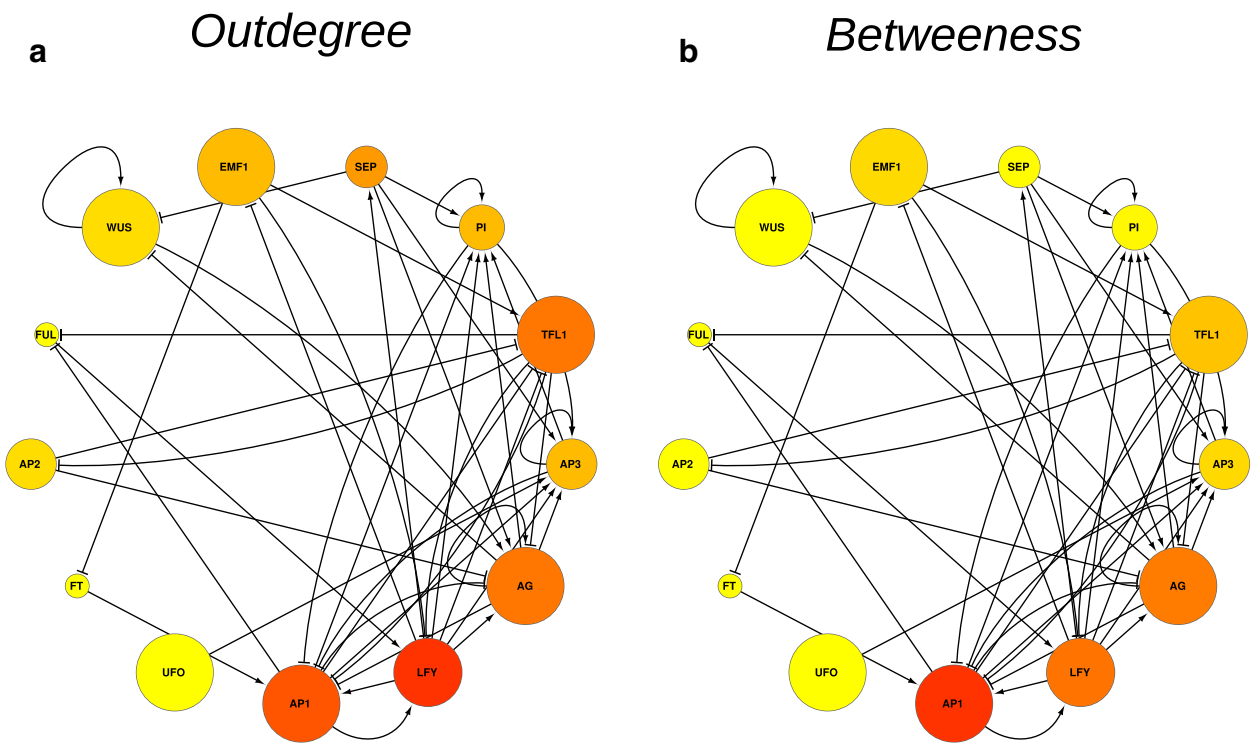


Figure 4 The FOS Gene regulatory network. The graph represents the mapping of the calculated *PT* values with the topological features out-degree **(a)** and betweenness **(b)** into the graph of the FOS-GRN. The size of the nodes represents the *PT* values. The topological features are represented by a graded yellow-red color scale with yellow (red) in the left (right) extreme.

graphically the associations by mapping the *PT* values and the topological features out-degree (Figure 4a) and betweenness (Figure 4b) into the graph of the FOS-GRN. The size of the nodes represents the *PT* values in the scale $[0, \ldots, 1]$. The topological feature is represented by a graded yellow-red color scale with yellow (red) in the left (right) extreme $[0, \ldots, 1]$.

Similarity in the propensity of interacting genes

To further test if there is an association between gene's topological features and their propensity to produce qualitative changes in the attractors, we performed the following analysis. Given the *PT* values for each gene, we asked if interacting genes within the FOS-GRN share more similar propensity within themselves than with non-interacting components. This pattern, if found, would suggest a close relationship between network architecture and such gene's dynamical property. Similar analyses have been proposed in network-based molecular evolutionary studies as a test for an association between network structure and evolutionary constraint [58,59]. In order to test whether this pattern is present in the FOS-GRN we calculated the average absolute difference (AAD) of the *PT* value between interacting components in the networks and used it as an statistic. An AAD of *PT* of 0.333 was calculated for the FOS-GRN. We then tested how likely is this value to be explained by change alone; specifically, we generated a null distribution by calculating AAD values in an ensemble of similar but random networks. We include the histogram of the corresponding statistic on an ensemble of 100,000 random networks with the same number of nodes and interactions in Additional file 1. Based on this data we estimated the probability of observing such a small value by calculating the fraction of random networks showing an AAD value $AAD \leq 0.333$ or greater. The resulting probability was 0.06.

Taken together these results: (1) a significant relationship between *PT* metric and the topological features of out-degree and betweeness, and (2) a marginally significant (p-value ~ 0.06) similar propensity within interacting genes; support the hypothesis that there is a close relationship between a gene's placement in the network, or its micro-topological position within a GRN, and its propensity to produce qualitative changes to the AL – at least in the case of the FOS-GRN. More general analyses for GRN with different topologies and architectures should be done.

Discussion

Recently, several authors have considered the restructuring of the state-space associated with a dynamical model of a GRN as an important aid for understanding

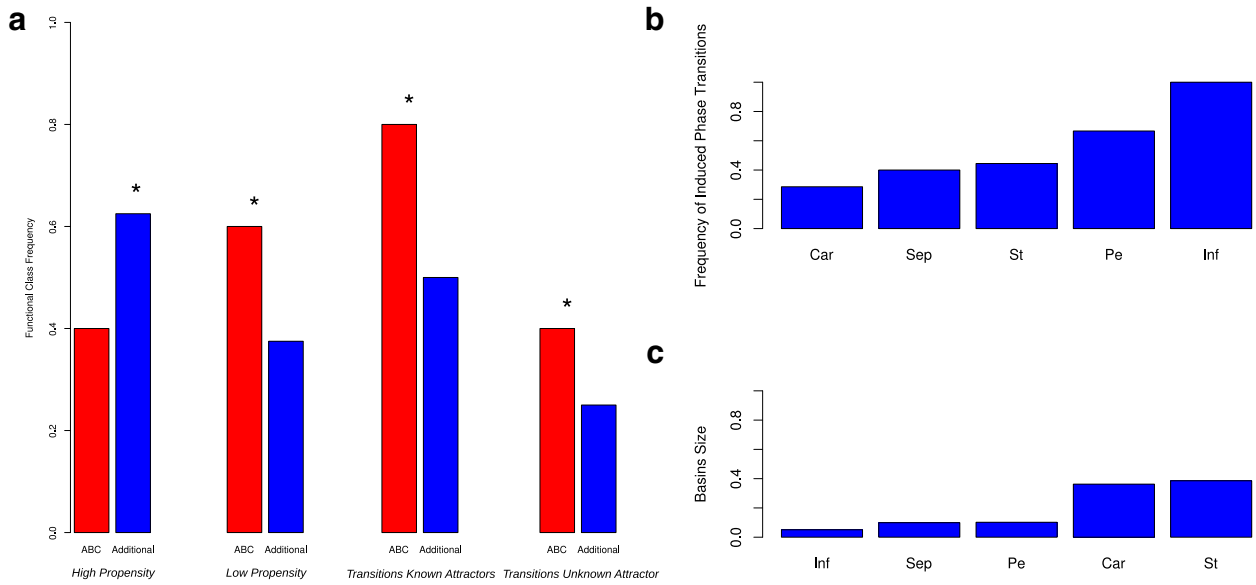


Figure 5 Genes propensity and functional class. **a)** The plot shows the gene frequency of each functional group (i.e., ABC, or Additional) within each gene class (i.e., High Propensity Genes, Low Propensity Genes, Genes Causing Transition Between Known Attractors, Genes Causing Transition Between Unknown Attractors). The star sign represents gene group over-representation as defined by a lower p-value relative to the other gene functional class calculated with a hypergeometric test (see Methods). **b)** The plot shows the calculated attractors propensity to undergo attractor transitions. **c)** Attractors basin size plot.

underlying mechanisms occurring during development an evolution [5,32,60-65]. A conclusion is emerging: the model of a landscape changing over time seems plausible as an explanation for fundamental features of morphogenesis and tissue formation [13]. In general, however, most work in this regard has been centered around either conceptual discussions or the dynamical analyses of small gene circuits. The exploration of such questions in larger, multi-attractor GRNs, that are grounded on experimental data and underlie realistic cases of cell differentiation, and in which the state-space presents a more complex structure, has largely been left behind. Here we present a modeling framework of general applicability as a first step for such type of exploration. For the sake of concreteness, we used as a model GRN the specific case of the FOS-GRN.

ODE-based models allows more flexible choice of network parameters reflecting, for example, different interaction strengths or inductive signals. Analyses of mathematical models of differentiation dynamics have shown that the considerations of such flexibility may be important to understand and control cell-fate choices (see, for example [5,9]). In the present case, given the hypothesis raised by some of the authors in [33] that differences in gene decay rates may potentially guide cell-fate decisions during flower development; we focus exclusively on the impact of relative gene decay rates in restructuring the AL, and thus we limit the scope of our conclusions. Additionally, the specific biological mechanisms driving such differential expression dynamics in vivo are not known. We speculate that signaling modules regulating responses to environmental cues may be directly connected to some of the components included in the GRN module analyzed here. In this direction, some of the authors have recently started to characterized such integrated GRNs considering the relevance of light sensing in flowering developmental choices [66]. Future work will test the effect of coupling such signaling modules with the GRN analyzed herein on the structure of the AL.

In the present case, when a given gene's decay rate is tuned and crosses a threshold, we observe qualitative changes in the AL's organization. We refer to the different patterns of organization as *phases*. The study of complex systems is, to a large extent, a search for the principles pervading self-organized, emergent phenomena and defining its potential phases [43,67]. Following this complex systems perspective, in this work we thus explored the phase changes in the AL that emerge from the dynamics of an experimentally grounded, complex GRN. Such transition phenomena are collective by nature and result from interactions taking place among the interacting genes of the GRN and not by any single gene alone. In any case, our exploration helped uncover a differential role of individual genes regarding their propensity to produce these induced *phase transitions*.

Given that the observed phase changes effectively correspond to qualitative changes of the AL in which one or more of the attractors (cell states) disappear, the result

would inevitably lead to an induced cell-fate decision. We focus on these latter attractors transitions. We must point out that in the present case we study the induced qualitative changes of the AL indirectly by systematically analyzing the local effects on each attractor of quantitative changes in gene decay rates. The relative stability of each attractor's basin is expected to be relevant in constraining transitions among attractors. This latter problem is the subject of current intense research and is more naturally approached by using stochastic models (see, for example [34,68]).

Differences in decay rates may also be interpreted as different time-scale regimes. Interestingly, a recent study stressed the relevance of time delays arising from multistep chemical reactions or cellular shape transformations [69]. Specifically, the authors argue in this reference that such feature is crucial in understanding cell differentiation, as it leads to novel states in epigenetic landscapes. In the present case, we indeed found that relatively different gene time-scale regimes produce qualitative changes to the otherwise static AL. Unlike the generic model presented by Mitra and collaborators [69], however, here we studied the dynamical behavior of specific genes which have been extensively characterized experimentally during decades of plant developmental genetics studies (see, for example [2]).

Most studies on the molecular basis of floral development focus on the eukaryotic MADS-box gene family, particularly floral homeotic genes such as AGAMOUS (AG), APETALA3 (AP3), PISTILLATA (PI), and several AGAMOUS-like genes [70]. Such genes are also the most important constituents of the ABC model for flower organ specification described above. Although based on extensive experimentation, the ABC genes have been characterized as having a prominent, functional role in cell fate and organ type specification during early flower development yielding homeotic transformations among floral organ when mutated; it was only a mechanistic view, the FOS-GRN dynamical model, which provided a sufficient explanation for the empirically observed ABC patterns – i.e., the combinatorial ABC code and the stable gene expression configurations observed during early flower development in *Arabidopsis* [2,11,19]. This model has been studied from different perspectives [24,33,41].

When testing the coherence of experimental data regarding the role of these molecular regulators under the framework of a GRN dynamical model certain questions arise. Why the ABC genes and not the other genes in the network display *homeotic* mutations when they are inactivated? Is there a relationship with this characterized biological (functional) property and its dynamical behavior within the FOS-GRN? What genes are more prone to have a stronger influence on the dynamical behavior of the whole system, and thus the phenotype, when perturbed or coupled with other circuits, signaling mechanisms, or processes outside the GRN module? Here we present a methodological framework for systematically testing the potential of specific genes when perturbed to produce qualitative changes to the underlying AL. By applying this approach to the FOS-GRN we uncover differences in the functional (dynamical) role of their genes. We speculate that such dynamical behavior might give information about which genes are most likely to be links with other circuits and processes.

A somewhat unexpected result is that the homeotic genes are less likely to produce attractor transitions in the AL by an induced higher decay rate, in comparison to other non-ABC genes in the network (see Methods). However, if we consider that ABC genes specify floral organ identity, a late process in early flower development, a higher robustness to non genetic perturbations such as changes in gene expression parameters is consistent with an increased stability of the cellular phenotypes as development proceeds. Indeed, when analyzing the propensity of the different attractors to undergo attractor transitions (see Methods) we found that the attractors corresponding to the flower cell-types show a lower propensity that the Inflorescence attractors (see below). On the other hand, we also found that in the cases where a phase transition induced by higher decay rates of ABC genes relative to the rates of other genes, the output is more likely to produce both an induced inter-attractor transition and to specify a novel attractor. This result aligns well with the empirical status of the ABC genes as *homeotic* genes, as it suggests that higher enough perturbations slowing gene function that approach a loss-of-function mutation, eliminate or produce specific cellular phenotypes, that correspond to changes of attractors, and thus homeotic alterations.

In Alvarez-Buylla and collaborators [24] some of the authors proposed a mechanistic explanation for the stereotypical temporal pattern of cell-fate specification during early flower development by means of noise-induced attractors transitions. In that study, however, it was shown that stochasticity alone was not able to explain a transition from the inflorescence to the flower meristems (attractors), an early, well-characterized event during flower development. Thus the authors speculate on the role of non-random inductive signals in the transition from cell fates in the inflorescence meristem to those in the flower meristem [24]. Our results suggest that this indeed could be the case, as a relatively higher decay rate of any of the genes expressed in the inflorescence attractors (TFL1, EMF1, UFO,WUS), with respect to the other genes, always produces a phase transition, and this transitions predominantly lead to flower organ attractors (see results). Thus, our model uncovered a potential mechanism which could be subjected to experimental validation. Namely, TFL1, EMF1, UFO, or WUS genes have a

relatively higher gene decay rate relative to flower specification genes during early flower development and within the inflorescence meristem. This feature in turn facilitates the inflorescence-flower transition when these genes are altered in their decay rates, thus suggesting that signals or pathways at play during the transition from inflorescence to flower meristem should interact or affect decay rates of these genes. In contrast, most functional studies concerning inflorescence to flower transition, have mostly focused on LFY and also on AP1 [71,72].

The distinction between molecular network structure and function is a core problem in systems biology. Dynamical GRN models enable a rigorous distinction between structure (topology) and function (dynamics). In a recent molecular evolutionary study also using the FOS-GRN, it was suggested that the dynamical functional role of genes within the network, and not just its connectivity, could play an important role in constraining evolution [59]. Such hypothesis implies a close relationship between network structure and function. Based on our operational definition of the gene functional role as the gene's propensity to produce AL attractor transitions, we asked if this property is associated with the gene's network topological features. We found that a significant correlation among these two. Our results thus support the hypothesis that for the FOS-GRN there is a close relationship between a gene's placement in the network and its propensity to produce attractor transitions in the AL. Likewise our results also provide partial support for the dynamical functional role of genes being important for constraining evolutionary changes.

Conclusions

In this contribution we present a methodology of general applicability as a first step for exploring the restructuring of the state-space associated with a dynamical multi-attractor GRN model. The framework consists on systematically exploring the propensity of single genes to produce qualitative changes in the AL as a result of changes in their parameters. Importantly, different GRN models and the effect of general inductive signals can be explored within the same framework. We showed how biological insights can be derived by applying the methodological framework to a single well-characterized and experimentally grounded GRN: the FOS-GRN. Future studies should explore if the results derived for this GRN can be generalized to GRN with contrasting typologies and architectures.

We systematically explored the effect of relative differences in gene decay rates on AL structure, and showed that by analyzing gene dynamics we can test if there are differences in the functional (dynamical) role among individual genes within the network, and that such differences correlate with biological observables. Specifically, (1) the dynamical behavior of ABC genes provide both robustness and flexibility in response to parameter perturbations, and are prone to both produce inter-attractor transitions and specify novel attractors; (2) It is less likely to induce attractor transitions as the course of development progresses; (3) non-random inductive signals may be at play in the transition from cell fates in the inflorescence meristem to those in the flower meristem; and (4) for the FOS-GRN there is a close relationship between a gene's placement in the network and its dynamical role. Taking together, our results suggest that there is a relationship between the impact of specific genes in the dynamics of the whole FOS-GRN, their biological function, and the observed hierarchy of differentiation events during early flower development.

Additional files

Additional file 1: Supporting Figures and Tables. The Additional file 1 includes the following information: Figure A1. Boolean FOS-GRN logical update rules. Figure A2. Attractors of the Wild-type Boolean FOS-GRN. Figure A3. ODEs model of the FOS-GRN. Figure A4. Attractors of the Wild-type ODEs FOS-GRN Model. Figure A5. Comparison of the Attractors and Basins Uncovered with the Boolean and ODEs FOS-GRN Models. Table A1. Table summarizing the result of all the bifurcation analyses. Figure A6. Histogram of the average absolute difference in PT values calculated from simulated networks values.

Additional file 2: Results of the Bifurcation (Phase transition) Analysis.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

ERAB coordinated the study and with the other authors established the overall logic and core questions to be addressed. All three authors conceived and planned the modeling approaches, JDV established many of the specific analyses to be done, recovered the information from the literature to establish the model, programmed and ran all the modeling and analyses. All authors participated in the interpretation of the results and analyses. JDV wrote most of the paper with help from ERAB and inputs from CV. All authors proofread the final version of the ms submitted. All authors read and approved the final manuscript.

Acknowledgements

Jose Davila-Velderrain acknowledges support from the graduate program Doctorado en Ciencias Biomédicas, Universidad Nacional Autónoma de México and CONACyT for financial support. This work is presented in partial fulfillment towards his doctoral degree in this Program. This work was supported by grants CONACyT:180380 (CV and ERAB), 180098 (ERAB); UNAM-DGAPA-PAPIIT: IN203113; IN204011; IN226510-3 and IN203814 to ERAB. We acknowledge Diana Romo for her help in many logistical tasks.

Author details

[1]Instituto de Ecología, Universidad Nacional Autónoma de México, Cd. Universitaria, 04510 México, D.F., México. [2]Centro de Ciencias de la Complejidad (C3), Universidad Nacional Autónoma de México, Cd. Universitaria, 04510 México, D.F., México. [3]Instituto de Física, Universidad Nacional Autónoma de México, Cd. Universitaria, 04510 México, D.F., México.

Received: 19 December 2014 Accepted: 22 April 2015
Published online: 13 May 2015

References

1. Alvarez-Buylla aER, Benítez M, Dávila EB, Chaos A, Espinosa-Soto C, Padilla-Longoria P. Gene regulatory network models for plant development. Curr Opin Plant Biol. 2007;10(1):83–91.
2. Alvarez-Buylla ER, Azpeitia E, Barrio R, Benítez M, Padilla-Longoria P. From abc genes to regulatory networks, epigenetic landscapes and flower morphogenesis: making biological sense of theoretical approaches. Seminars Cell Dev Biol. 2010;21(1):108–17.
3. Furusawa C, Kaneko K. A dynamical-systems view of stem cell biology. Science. 2012;338(6104):215–7.
4. Huang S, Kauffman S. Complex gene regulatory networks-from structure to biological observables: cell fate determination In: RA M, editor. Encyclopedia of Complexity and Systems Science. New York: Springer; 2009. p. 1180–293.
5. Huang S, Guo Y-P, May G, Enver T. Bifurcation dynamics in lineage-commitment in bipotent progenitor cells. Dev Biol. 2007;305(2): 695–713.
6. Huang S. Cell lineage determination in state space: a systems view brings flexibility to dogmatic canonical rules. PLoS Biol. 2010;8(5):1000380.
7. Andrecut M, Halley JD, Winkler DA, Huang S. A general model for binary cell fate decision gene circuits with degeneracy: indeterminacy and switch behavior in the absence of cooperativity. PloS One. 2011;6(5): 19358.
8. Zhou JX, Brusch L, Huang S. Predicting pancreas cell fate decisions and reprogramming with a hierarchical multi-attractor model. PloS One. 2011;6(3):14752.
9. Li C, Wang J. Quantifying cell fate decisions for differentiation and reprogramming of a human stem cell network: landscape and biological paths. PLoS Comput Biol. 2013;9(8):1003165.
10. Jaeger J, Sharpe J. On the concept of mechanism in development. In: Towards a Theory of Development. Oxford: Oxford University Press; 2014. p. 56.
11. Espinosa-Soto C, Padilla-Longoria P, Alvarez-Buylla ER. A gene regulatory network model for cell-fate determination during arabidopsis thaliana flower development that is robust and recovers experimental gene expression profiles. Plant Cell Online. 2004;16(11):2923–39.
12. Von Dassow G, Meir E, Munro EM, Odell GM. The segment polarity network is a robust developmental module. Nature. 2000;406(6792): 188–92.
13. Kaneko K. Characterization of stem cells and cancer cells on the basis of gene expression profile stability, plasticity, and robustness. Bioessays. 2011;33(6):403–13.
14. Kauffman SA. Metabolic stability and epigenesis in randomly constructed genetic nets. J Theor Biol. 1969;22(3):437–67.
15. Kauffman SA. The Origins of Order: Self-organization and Selection in Evolution. New York: Oxford university press; 1993.
16. Huang S, Eichler G, Bar-Yam Y, Ingber DE. Cell fates as high-dimensional attractor states of a complex gene regulatory network. Phys Rev Lett. 2005;94(12):128701.
17. Chang HH, Hemberg M, Barahona M, Ingber DE, Huang S. Transcriptome-wide noise controls lineage choice in mammalian progenitor cells. Nature. 2008;453(7194):544–7.
18. Albert R, Othmer HG. The topology of the regulatory interactions predicts the expression pattern of the segment polarity genes in drosophila melanogaster. J Theor Biol. 2003;223(1):1–18.
19. Mendoza L, Alvarez-Buylla ER. Dynamics of the genetic regulatory network for arabidopsis thaliana flower morphogenesis. J Theor Biol. 1998;193(2):307–19.
20. Albert I, Thakar J, Li S, Zhang R, Albert R. Boolean network simulations for life scientists. Source Code Biol Med. 2008;3(1):1–8.
21. Azpeitia E, Davila-Velderrain J, Villarreal C, Alvarez-Buylla ER. Gene regulatory network models for floral organ determination. In: Flower Development. New York: Springer; 2014. p. 441–69.
22. Waddington CH. The Strategy of Genes. London: George Allen & Unwin, Ltd.; 1957.
23. Siegal ML, Bergman A. Waddington's canalization revisited: developmental stability and evolution. Proc Nat Acad Sci. 2002;99(16): 10528–32.
24. Álvarez-Buylla ER, Chaos Á, Aldana M, Benítez M, Cortes-Poza Y, Espinosa-Soto C, et al. Floral morphogenesis: stochastic explorations of a gene network epigenetic landscape. Plos One. 2008;3(11):3626.
25. Wang J, Zhang K, Xu L, Wang E. Quantifying the waddington landscape and biological paths for development and differentiation. Proc Nat Acad Sci. 2011;108(20):8257–62.
26. Enver T, Pera M, Peterson C, Andrews PW. Stem cell states, fates, and the rules of attraction. Cell Stem Cell. 2009;4(5):387–97.
27. Fagan MB. Waddington redux: models and explanation in stem cell and systems biology. Biol Philosophy. 2012;27(2):179–213.
28. Ladewig J, Koch P, Brüstle O. Leveling waddington: the emergence of direct programming and the loss of cell fate hierarchies. Nat Rev Mol Cell Biol. 2013;14(4):225–36.
29. Huang S. The molecular and mathematical basis of waddington's epigenetic landscape: A framework for post-darwinian biology? Bioessays. 2012;34(2):149–57.
30. Davila-Velderrain J, Martinez-Garcia J, Alvarez-Buylla ER. Modeling the epigenetic attractors landscape: towards a post-genomic mechanistic understanding of development. Front Genet. 2015;6:160.
31. Zhou JX, Qiu X, d'Herouel AF, Huang S. Discrete gene network models for understanding multicellularity and cell reprogramming: From network structure to attractor landscapes landscape. In: Computational Systems Biology. CA: Elsevier; 2014. p. 241–76.
32. Davila-Velderrain J, Alvarez-Buylla ER. Bridging genotype and phenotype. In: Frontiers in Ecology, Evolution and Complexity. Coplt ArXives; 2014. p. 144-154.
33. Villarreal C, Padilla-Longoria P, Alvarez-Buylla ER. General theory of genotype to phenotype mapping: derivation of epigenetic landscapes from n-node complex gene regulatory networks. Phys Rev Lett. 2012;109(11):118102.
34. Zhou JX, Aliyu M, Aurell E, Huang S. Quasi-potential landscape in complex multi-stable systems. J R Soc Interface. 2012;9(77):3539–53.
35. Choi M, Shi J, Jung SH, Chen X, Cho K-H. Attractor landscape analysis reveals feedback loops in the p53 network that control the cellular response to dna damage. Sci Signaling. 2012;5(251):83.
36. Wang P, Song C, Zhang H, Wu Z, Tian X-J, Xing J. Epigenetic state network approach for describing cell phenotypic transitions. Interface Focus. 2014;4(3):20130068.
37. Lu M, Onuchic J, Ben-Jacob E. Construction of an effective landscape for multistate genetic switches. Phys Rev Lett. 2014;113(7):078102.
38. Fujimoto K, Ishihara S, Kaneko K. Network evolution of body plans. PLoS One. 2008;3(7):2772.
39. Suzuki N, Furusawa C, Kaneko K. Oscillatory protein expression dynamics endows stem cells with robust differentiation potential. PLoS One. 2011;6(11):27232.
40. Cotterell J, Sharpe J. Mechanistic explanations for restricted evolutionary paths that emerge from gene regulatory networks. PloS one. 2013;8(4): 61178.
41. Sanchez-Corrales Y-E, Alvarez-Buylla ER, Mendoza L. The arabidopsis thaliana flower organ specification gene regulatory network determines a robust differentiation process. J Theor Biol. 2010;264(3):971–83.
42. Coen ES, Meyerowitz EM. The war of the whorls: genetic interactions controlling flower development. Nature. 1991;353(6339):31–7.
43. Sole R. Phase Transitions. New Jersey: Princeton U. Press; 2011.
44. Seydel R. Practical Bifurcation and Stability Analysis. New York: Springer; 2010.
45. Müssel C, Hopfensitz M, Kestler HA. Boolnet–an r package for generation, reconstruction and analysis of boolean networks. Bioinformatics. 2010;26(10):1378–80.
46. Glass L. Classification of biological networks by their qualitative dynamics. J Theor Biol. 1975;54(1):85–107.
47. Mendoza L, Xenarios I. A method for the generation of standardized qualitative dynamical systems of regulatory networks. Theor Biol Med Modell. 2006;3(1):13.
48. Mangan S, Alon U. Structure and function of the feed-forward loop network motif. Proc Nat Acad Sci. 2003;100(21):11980–5.
49. Lu M, Jolly MK, Gomoto R, Huang B, Onuchic J, Ben-Jacob E. Tristability in cancer-associated microrna-tf chimera toggle switch. J Phys Chem B. 2013;117(42):13164–74.
50. Soetaert K, Petzoldt T, Setzer RW. Solving differential equations in r: Package desolve. J Stat Software. 2010;33(9):1–25.
51. Soetaert K, Cash J, Mazzia F. Solving Differential Equations in R. New York: Springer; 2012.

52. Shmulevich I, Kauffman SA, Aldana M. Eukaryotic cells are dynamically ordered or critical but not chaotic. Proc Nat Acad Sci USA. 2005;102(38): 13439–44.
53. James G, Witten D, Hastie T, Tibshirani R. An Introduction to Statistical Learning. New York: Springer; 2013.
54. Csardi G, Nepusz T. The igraph software package for complex network research. InterJournal Complex Syst. 2006;1695(5):1-9.
55. Barrio R. Á, Hernandez-Machado A, Varea C, Romero-Arias JR, Alvarez-Buylla E. Flower development as an interplay between dynamical physical fields and genetic networks. PloS one. 2010;5(10):13523.
56. Everitt B, Hothorn T. An Introduction to Applied Multivariate Analysis with R. New York: Springer; 2011.
57. Yu H, Huang J, Zhang W, Han J-DJ. Network analysis to interpret complex phenotypes. In: Applied Statistics for Network Biology: Methods in Systems Biology. Germany: Wiley Online Library; 2011. p. 1–12.
58. Alvarez-Ponce D, Aguadé M, Rozas J. Network-level molecular evolutionary analysis of the insulin/tor signal transduction pathway across 12 drosophila genomes. Genome Res. 2009;19(2):234–42.
59. Davila-Velderrain J, Servin-Marquez A, Alvarez-Buylla ER. Molecular evolution constraints in the floral organ specification gene regulatory network module across 18 angiosperm genomes. Mol Biol Evol. 2014;31(3):560–73.
60. Pujadas E, Feinberg AP. Regulated noise in the epigenetic landscape of development and disease. Cell. 2012;148(6):1123–31.
61. Ferrell Jr JE. Bistability, bifurcations, and waddington's epigenetic landscape. Curr Biol. 2012;22(11):458–66.
62. Wang J, Xu L, Wang E, Huang S. The potential landscape of genetic circuits imposes the arrow of time in stem cell differentiation. Biophys J. 2010;99(1):29–39.
63. Jaeger J, Crombach A. Life's attractors. In: Evolutionary Systems Biology. New York: Springer; 2012. p. 93–119.
64. Jaeger J, Irons D, Monk N. The inheritance of process: a dynamical systems approach. J Environ Zool Part B: Mol Dev Evol. 2012;318(8): 591–612.
65. Verd B, Crombach A, Jaeger J. Classification of transient behaviours in a time-dependent toggle switch model. BMC Syst Biol. 2014;8(1):43.
66. Pérez-Ruiz RV, García-Ponce B, Marsch-Martínez N, Ugartechea-Chirino Y, Villajuana-Bonequi M, de Folter S, et al. XAANTAL2 (AGL14) is an important component of the complex gene regulatory network that underlies arabidopsis shoot apical meristem transitions. Mol Plant. 2015. doi:10.1016/j.molp.2015.01.017.
67. Haken H. Synergetics. New York: Springer; 1977.
68. Ge H, Qian H. Landscapes of non-gradient dynamics without detailed balance: Stable limit cycles and multiple attractors. Chaos: Interdisciplinary J Nonlinear Sci. 2012;22(2):023140.
69. Mitra MK, Taylor PR, Hutchison CJ, McLeish T, Chakrabarti B. Delayed self-regulation and time-dependent chemical drive leads to novel states in epigenetic landscapes. J R Soc Interface. 2014;11(100):20140706.
70. Lawton-Rauh AL, Alvarez-Buylla ER, Purugganan MD. Molecular evolution of flower development. Trends Ecol Evol. 2000;15(4):144–9.
71. Mandel MA, Yanofsky MF. A gene triggering flower formation in arabidopsis. Nature. 1995;377(6549):522–4.
72. Benlloch R, Berbel A, Serrano-Mislata A, Madueño F. Floral initiation and inflorescence architecture: a comparative view. Ann Bot. 2007;100(3): 659–76.

Dynamic network and epigenetic landscape model of a regulatory core underlying spontaneous immortalization and epithelial carcinogenesis

Méndez-López LF[1,2,†], Davila-Velderrain J[1,2,†], Enríquez-Olguín C[3], Martinez-Garcia JC[3,*], Alvarez-Buylla ER[1,2,*]

1 Instituto de Ecología, Universidad Nacional Autónoma de México, Cd. Universitaria, México, D.F. 04510, México

2 Centro de Ciencias de la Complejidad (C3), Universidad Nacional Autónoma de México, Cd. Universitaria, México, D.F. 04510, México

3 Departamento de Control Automático, Cinvestav-IPN, A. P. 14-740, 07300 México, DF, México

∗ **Corresponding authors: juancarlos_martinez-garcia@conciliencia.org, eabuylla@gmail.com**

† **These authors contributed equally to this work**

Abstract

Tumorigenic transformation of human epithelial cells *in vitro* has been described experimentally as the potential result of a process known as *spontaneous immortalization*. In this process a generic series of cell–state transitions occur in which normal epithelial cells acquire a senescent state, later surpassed to attain a mesenchymal state and finally a mesenchymal stem–like phenotype, with a potential tumorigenic behavior. In this paper we integrate published data on the molecular components and interactions that have been described as key regulators of such cell states and transitions. Such large network, that is provided, is then reduced with the aim of recovering a minimal regulatory core incorporating the necessary and sufficient restrictions to recover the observed cell states and their generic progression patterns in epithelial–mesenchymal transition. Data is formalized into logical regulatory rules that govern the dynamics of each of the networks components as a function of the states of its regulators. The proposed core gene regulatory network attains only three steady–state gene expression configurations that correspond to the profiles characteristic of normal epithelial, senescent, and mesenchymal stem–like cells. Interestingly, epigenetic analyses of the uncovered network shows that it also recovers the generic time–ordered transitions documented during tumorigenic transformation *in vitro* of epithelial cells, and which strongly correlate with the patterns observed during the progressive pathological description of epithelial carcinogenesis *in vivo*.

Introduction

Nearly 84% of cancers diagnosed in human adults are carcinomas (i.e., cancer of epithelial origin), and their emergence is strongly associated with both an underlying chronic inflammatory process and with aging [1]. The precise role and the contribution of these two processes to the origin, progression, and detected clinic behavior of epithelial cancers remains elusive, however. The current general assumption is that aging and inflammation increase the chance of accumulating somatic mutations, and this genetic instability ultimately leads to carcinoma. However, this view does not offer a logical or mechanistic explanation for well–documented observations. For example: (1) cancer cells show morphological and transcriptional convergences despite their diverse origin, (2) carcinogenesis recapitulates embryonic processes, (3) cancer behavior can be acquired in the absence of mutations through trans– or dedifferentiation, and (4) cancer cells can be "normalized" by several experimental means [2–5]. Moreover, it is well–known that different carcinomas share the same cellular processes and histological stages or progression patterns, as well as robust associations with lifestyle factors [6]. These empirical observations

suggests that, in analogy to normal development, the human genome is associated with an underlying robust mechanism restricting cell states and temporal progression patterns that are characteristic of epithelial carcinogenesis. In accordance with this view, other researchers have previously proposed that cancer can be considered a developmental disease [7, 8].

In systems biology it is common to understand both cell differentiation and development in terms of dynamical systems theory. In this framework, the genome of a cell is directly mapped into a global and multi–stable gene regulatory network (GRN) whose dynamics yields several (quasi)stationary and stable distinct phenotypic cellular states [9–14]. That is, the same genome robustly generates multiple discrete cellular phenotypes through developmental dynamics [12, 15, 16]. These stable phenotypic states are called *attractors* and correspond to configurations of gene or protein activation states that underlie the cellular fates or phenotypes – *i.e.*, which thus constitute biological observables. Therefore, developmental processes – cellular differentiation events in particular – are formalized in temporal terms as attractor's (*i.e.*, cell states) transitions. Here we adopt such approach to study the cell states attained and the time–ordered transitions observed during the tumorigenic transformation of epithelial cells cultured *in vitro* that surpass a senescent state; a process known as *spontaneous immortalization*.

Experimental findings in molecular and cell biology of cancer research have revealed that it is possible to recover cells with cancer–like phenotypes through some specific cellular transitions. This has been shown particularly in carcinomas [3, 17–19]. By a cellular transition we refer to a differentiation event in which a certain cell acquires a discretely different cellular phenotype. For example, the process called epithelial–mesenchymal transition (EMT) comprises a stereotypical cell state transition in which epithelial cells exposed, for example, to cytokines, are induced to undergo a discrete phenotypic change acquiring a mesenchymal phenotype [17,19]. Interestingly, through inflammation–induced EMT epithelial cells surpass senescence, and undergo spontaneous immortalization. Cells that emerge from this process manifest mesenchymal stem–like properties and are capable of developing cancer in murine models [3,18]. Furthermore, these cells are difficult to distinguish phenotypically and in terms of the transcription factors that they express from either the so–called cancer stem cells (also known as tumor initiating cells) or from embryonic stem cells [20, 21].

In the present work, we hypothesize that a generic series of cell state transitions widely observed and robustly induced by inflammation in cell cultures during spontaneous immortalization naturally result from the self–organized behavior emerging from an underlying intracellular GRN. During this process, normal epithelial cells first acquire a senescent state, to finally attain a mesenchymal stem–like cellular state with a potential tumorigenic behavior. We speculate that tissue–level conditions associated with a bad prognosis, such as a pro–inflammatory milieu, may increase the rate of occurrence of these same transitions *in vivo* promoting as a result the emergence and progression of epithelial cancer.

In an attempt to provide mechanistic insights into the regulation of the aforementioned observed cell–fates specification, as well as the time–ordered cell–state transitions, we propose here a cellular level GRN model that integrates the available experimental data concerning the main molecular components and interactions related to the emergence and progression of carcinomas. We propose a large GRN of 41 nodes that integrates cellular processes thoroughly studied experimentally, but which have not been integrated before into a single GRN. Specifically, the large GRN model includes key molecular regulators that: (1) characterize the cellular phenotypes of epithelial, mesenchymal, and senescent cells; (2) are involved in the induction of the cellular processes of replicative senescence, cellular inflammation, and EMT; and (3) characterize the phenotypic changes undergone by cells emerging from these processes (*i.e.* mesenchymal stem–like cells). To obtain a minimal regulatory core for further dynamical analyses we formally reduced the large GRN. We show that the proposed regulatory core module displays an orchestrating robust behavior akin to that seen in other developmental regulatory modules previously characterized with similar formal approaches (see, for example [9, 10, 22, 23]). Specifically, by proposing logical functions grounded on experimental data for this regulatory core module and by analyzing its behavior following conventional Boolean GRN dynamical approaches, we show that the uncovered minimal GRN converges only to three

attractors. The uncovered states correspond to the expected gene expression configurations that have been observed for normal epithelial, senescent and stem–like mesenchymal cellular fates. Additionally, we also explore the GRN Epigenetic Landscape using a stochastic version of the model (following: [24, 25]) in order to address if the proposed GRN also restricts or underlies the generic temporal sequence with which cell states occur in cell cultures and which correlate with observed patterns of cell–type enrichment during pathological descriptions of carcinoma progression.

Results

Gene Regulatory Network Construction

Following a bottom–up and an expert knowledge approach we propose a set of cellular dynamical processes ubiquitous to epithelial carcinogenesis, namely: replicative cellular senescence, inflammation, and epithelial–mesenchymal transition (EMT). The cellular phenotypes epithelial, senescent, and mesenchymal cell–types – as well as a mesenchymal embryonic–like state; have been largely characterized as biological observables involved in such processes. We provide further definitions of these – and associated – phenotypes and processes in our complementary Text S1. We take this information as a methodological basis to integrate a generic dynamical network model of epithelial carcinogenesis. As a first step in network integration, based on an extensive literature search (see Methods and Text S1), we assembled a set of transcription factors (TFs) and additional molecules involved in the establishment and regulation of these cellular states and processes. Subsequently, we manually retrieved documented regulatory interactions among the molecules, considering only those supported by experimental evidence. For a detailed description of the published information for each interaction proposed see Text S1. The constructed large GRN is shown in Figure 1 (see Methods). TFs are represented in graphical terms by squares and the rest of the molecules by circles. The identified large network consists of 41 nodes and 97 interactions; it includes 12 TFs which can be considered as key regulators of the processes under consideration. Colors indicate the association that each node hold with specific cellular phenotypes or processes being considered: epithelial (green), mesenchymal (orange), inflammation (red), senescence and DNA damage (blue), cell–cycle (purple), and polycomb complex (yellow).

The Proposed Network is Enriched with Cancer Pathways

In order to provide additional partial support for the association of the bio-molecular set of regulatory interactions that we have manually curated based on published data with the processes under consideration, as well as with carcinoma, we performed a network–based gene set enrichment analysis (GSEA) (see Methods). Among the 13 pathways or processes reported as significant when taking the KEGG database as a reference, 9 (69%) correspond to cancer pathways, namely: *Bladder cancer, Chronic myeloid leukemia, Pancreatic cancer, Glioma, Non–small cell lung cancer, Melanoma, Small cell lung cancer, Prostate cancer, and Thyroid cancer* – note that 6 (66.6%) of these correspond to carcinomas. On the other hand, when taking the GO Biological Process database as reference, among the significant results we found: *replicative senescence, cellular senescence, cell aging, activation of NF–κB–inducing kinase activity, determination of adult life span, epithelial cell differentiation, and positive regulation of NF–κB transcription factor activity* (see Table 1). Using network topological gene set analysis (see Methods) we found that, in addition to pathway enrichment, the topological signature of the molecules in the proposed network also shows a topological signature that is similar to the one shown by reference cancer pathways included in the KEGG database (see Figure S1). These results provide partial support for the proposed molecular players: given the current state of knowledge according to annotated databases, the set of molecules manually included in the proposed large network seems to be representative of the cellular phenotypes and processes considered as prior biological knowledge in our model. In addition,

the molecular components included in the proposed large network are tightly associated with reference pathways of epithelial cancers.

A Core Regulatory Network Module Underlying Spontaneous Immortalization

We performed a knowledge–based network reduction of the large GRN in Figure 1 in order to derive a smaller, core GRN module for which both a topology and architecture with fully defined logical functions could be established, and which could also be analyzed as a dynamical system (see Methods). In addition, such regulatory core should comprise the necessary and sufficient set of nodes and interactions that integrate the processes involved in the large network and that could explain, at least in part, the restricted set of the cell–states and time–ordered transitions among them during spontaneous immortalization and epithelial cancer emergence/progression. We were able to define a set of molecular species whose regulatory hierarchy, activity, and expression define the identity of the phenotypes of epithelial, mesenchymal, and senescent cells. We also converged to, and included, main regulators of replicative cellular senescence, inflammation–induced EMT, and determinants of an induced mesenchymal stem–like phenotype. Hence, after reduction we obtained a core GRN consisting of only 9 nodes: *ESE–2, Snai2, NF–κB, E2F, p53, p16, Rb, Cyclin, and Telomerase*. Figure 1b shows the proposed core regulatory module (colored nodes) in the context of the larger proposed network. For details on how these 9 nodes were selected over the rest of the nodes see Text S1. In what follows we present a brief description of the nodes included in the reduced GRN, as well as some of the key molecular mechanism encoded in the regulatory logic. Although many of the nodes that are included in this regulatory core module have been thoroughly studied experimentally and in terms of their involvement in different types of cancer, the architecture and topology of the proposed regulatory core module is novel.

ESE–2 represents the activity of the TFs ESE–1, ESE–2, and ESE–3 (also known as ESX, E74–like factor 5, and EHF; respectively) – for a table with synonyms Table E1 in supplementary file. These proteins belong to the subgroup ESE (*i.e.* epithelium–specific) of the TF family ETS. ESE–2 promotes its own expression and the expression of the other ESE TFs [26–28]. On the other hand, ESE–2 represses Snai2 – one of the main EMT promoting TFs – expression by direct interaction with its promoter region [29].

p16 represents the activity of the INK4b–ARF–INK4a locus, which encodes for the proteins p16 and p14. Cellular senescence is molecularly characterized by the expression of the proteins p16 and p53 [30]. p16 indirectly inhibits E2F by inhibiting cyclins CDK 2,4 and 6, which in turn inhibit Rb [31, 32]. On the other hand, the INK4b–ARF–INK4a – and thus p16 – is regulated by the activity of Polycomb–group proteins by means of promoter hypermethylation [33].

p53 represents the protein with the same name. The shortage of telomeric DNA seems to be recognized as DNA damage promoting the activation of p53. In senescence, the activity of p16 and p53 over Rb, E2F and Cyclins invariably arrests the cell–cycle in the phase G1/G2 [34, 35].

Rb represents the cell–cycle regulator with the same name. Rb prevents cycle progression by forming a complex with the TF E2F [36].

E2F represents the TF with the same name. E2F regulates critical genes for adequate cell–cycle progression.

Cyclin represents the activity of the complex Cyclin–dependent kinases (CDKs) known to inactivate Rb by phosphorylation. The latter, in turn, promotes the activity of E2F and cell–cycle progression [37].

NF-κB represents cellular inflammation by the activity of the TF NF–κB. Accordingly, with this node we also represent the effect of the cytokines transforming growth factor–beta (TGF–β), interleukin–6 (IL–6), and tumor necrosis factor alpha (TNF–α). These three factors converge in the activation of NF–κB by phosphorylating the inhibitor IκB [38, 39].

TELasa represents the enzyme telomerase. This enzyme is responsible for the *de novo* synthesis of telomeres. Most human cell–types do not express telomerase; however, it is expressed on immortalized epithelial cells, and it is thought to be responsible for telomere extension in tumors [40].

Snai2 this node includes the activity of the main TFs known to be directly associated with EMT regulation, namely: Snai2 (Slug), Snail, Twist1, Twist2, ZEB1, ZEB2, and FOXC2. These TFs repress (induce) the expression of genes specific to epithelial (mesenchymal) cells [41, 42]. It has been proposed that there is a regulatory hierarchy driving EMT in which Snail activates Snai2, Twist, Zeb, and FOXC2. The latter, in turn, regulates Snail and Snai2 in a positive manner [41, 43–45]. Regardless of a hierarchical interpretation, it is well–documented that these TFs maintain the mesenchymal phenotype in a coordinated fashion, showing co–expression patterns and regulatory crosstalk [44, 45]. It has been suggested that among these TFs, Snai2 may be the strongest suppressor of the epithelial phenotype [46]. However, we decided to represent the collective regulatory activity of the mesenchymal TFs using Snai2 based on the recent experimental demonstration of an antagonistic relation between Snai2 and ESE–2. Specifically, *in vitro* and *in vivo* studies showed that ESE–2 regulates the transcription of Snai2 [29].

According to our model reduction methodology, literature search, careful manual curation, and network–based enrichment analysis; we propose that the derived core GRN module (see Figure 2) includes a molecular set which is both *necessary and sufficient* to specify the identity of the aforementioned cellular phenotypes and to represent the main intracellular regulatory events driving spontaneous immortalization in a robust manner. We test our proposal by building and analyzing a mechanistic GRN dynamical model (see below).

Recovered Attractors of the Core GRN Module Correspond to Configurations that Characterize Expected Cellular Phenotypes

Based on the experimental data concerning the expression patterns of the genes incorporated in the proposed core GRN model in Figure 2 we assembled a table with a Boolean format of the state configurations expected to be recovered with the proposed GRN dynamical model. We refer to this configurations as the *"expected attractors"* – these correspond to the empirically observed genetic configurations. Furthermore, we integrated and formalized the experimental data concerning the interactions among the GRN nodes using Boolean logical functions that will rule the Boolean GRN dynamics and comprise the architecture of the proposed GRN. The set of formulated rules underlying the regulatory events is shown in Text S1 – each logical rule is presented both as a logical preposition and as a truth table. Using the set of nodes and their corresponding logical rules we completely define a mechanistic dynamical GRN model [47]. The exhaustive computer–based simulation analysis of this model (see Methods) recovered three fixed–point attractors. Interestingly, the recovered attractors showed perfect correspondence with the expected attractors representing cellular phenotypes (see Table 2). The three recovered attractors correspond to the expected epithelial, senescent, and mesenchymal stem–like phenotypes :

The normal epithelial cell phenotype is represented by the attractor with ESE–2, E2F, Cyclin and NF–κB activity. ESE–2 is an epithelial–specific TF which regulates a large number of genes specific to epithelial cells [48, 49]. NF–κB shows ubiquitous expression through the different types of human cells; however, it is also positively regulated by TFs of the ESE family (*i.e.* ESE–2) [50]. Moreover, under

inflammatory conditions the activity of NF–κB is enhanced [51, 52]. On the other hand, E2F and Cyclin represent core regulators of cell–cycle entrance, and thus specify proliferative capability [53, 54].

The senescent cell phenotype is represented by the attractor with ESE–2, Rb, p16, p53, and NF–κB activity. Its biological counterpart would be an epithelial cell induced to replicative senescence, given (1) that it is expected to repress E2F [48]; and (2) that Rb, p16, p53, and NF–κB are the molecular biomarkers of cellular senescence [55].

Messenchymal Stem-like phenotype In the model proposed here, the attractor whose configuration shows Snai2, Cyclin, NF–κB, and Telomerase activity – and inactivity of ESE–2, p16, Rb, p53, and E2F – would correspond to a mesenchymal stem–like phenotype with tumorigenic potential (see discussion below).

The correspondence between the recovered attractors and the expected cellular phenotypes strongly suggests that the proposed nine–node core GRN indeed constitutes a regulatory module that is robust to initial conditions and that comprises a set of necessary and sufficient components and interactions to restrict the system to converge to the cellular phenotypes observed during spontaneous immortalization.

Validation of the Uncovered Core Regulatory Module: Loss and gain–of–function Mutant and Robustness Analyses

In order to validate the Boolean GRN dynamics we tested if the same GRN module is able to recover observed attractors in loss and gain of function mutants. We simulated such mutants analogous to experimental observations reported in the literature. Specifically, we simulated loss– and gain–of–function mutations of ESE–2, Snai2, and p16 that have been reported in the literature. When simulating ESE–2 gain of function (by setting the expression state for this node permanently to "1" in the simulations), the GRN model recovers three attractors corresponding to three different phenotypes which have been experimentally described and are associated with ESE–2 over–expression: an epithelial senescent cell [56], a normal epithelial cell [29], and a metastable state with proliferative phenotype [57]. In the case of ESE–2 loss–of–function (simulated by setting the expression state of this node to "0" permanently), the model recovers an attractor corresponding to a mesenchymal phenotype, which is also consistent with observations [29]. For Snai2, gain–of–function simulation recovers one attractor corresponding to mesenchymal stem–like phenotype, which is consistent with observations from ectopic over–expression experiments of mesenchymal TFs [18, 58, 59]. Snai2 loss–of–function simulation, on the other hand, recovered two attractors corresponding to normal and senescent epithelial phenotypes, which is also consistent with observations [29, 60]. Finally, gain–of–function simulation of p16 recovered two attractors; one associated with a mesenchymal stem–like but incompletely senescent (due to the lack of p53) phenotype; the other corresponding to an epithelial senescent phenotype. The first prediction is consistent with the status of immortal and apoptosis–resistant shown by mesenchymal stem–like cells, as well as with the capability of mesenchymal TFs to abrogate senescence [61]. The second attractor is consistent with the potential for replicative senescence of epithelial cells. p16 loss–of–function simulation recovers two attractors corresponding to an epithelial cell and a mesenchymal stem–like cell. This prediction is consistent with the observed biological conditions for both phenotypes, where p16 is commonly repressed by polycomb proteins [62]. The recovered attractors in mutant conditions are shown in Figure S2 in supplementary file.

It is important to note that, given that the uncovered regulatory module uncovered here is the result of a model reduction methodology where we permissively chose to represent multiple molecular species by the activity of some of the nodes, a direct interpretation of mutant simulations is not straightforward. Consequently, care should be taken when interpreting the results of the simulations or making predictions

of mutant phenotypes yet to be experimentally tested and further explored in the context of the larger GRN in Figure 1, which is the focus of an ongoing study. With this in mind, instead of simulating additional mutant conditions, we further validated the dynamical GRN model by testing its robustness to perturbations of the logical rules. Specifically, we tested the robustness of the predicted attractors by generating a large set of perturbed networks (*e.g*, 10,000), calculating their respective attractors, and then counting the occurrences of the original attractors within the perturbed set. We generated each perturbed network by choosing a function of the network at random and flipping a single bit in this function [63]. We performed four complementary *in silico* based experiments following this general robustness analysis. First, we estimated the fraction of occurrences of the three original attractors (*i.e.*, their robustness). Then, we repeated the experiment three times, but each time estimating the robustness of each individual attractor. For these four experiments we estimated a robustness (*i.e.*, fraction of times) of 0.7439, 0.905, 0.923, and 0.902, respectively. Hence, out of 10,000 random networks generated by *in silico* perturbations to the logical rules, a major fraction recovered the original attractors; as it is expected for a developmental (core) regulatory module that is robust both to transient (initial) and genetic perturbations [10]. This result supports the view that the core GRN uncovered here is indeed a regulatory network module driving developmental dynamics. It also constitutes a mechanistic explanation (for definitions, see [47]) to the generic cell phenotypes observed during spontaneous immortalization *in vitro* and which correlate with the cellular description of carcinoma progression *in vivo* (see below).

Attractor Time–Ordered Transitions: Epigenetic Landscape of the Uncovered GRN Core Module

During the tumorigenic transformation of epithelial cells in culture, a generic time–ordered series of cell state transitions is observed and robustly induced by inflammation [3, 18]. Normal epithelial cell become senescent cells, which afterwards overcome this latter state acquiring a final mesenchymal stem–like phenotype. Interestingly, during the progressive pathological description of epithelial carcinomas *in vivo* the temporal pattern with which each of these different cell phenotypes enriches the tissue seems to be tightly ordered and is also generic to all types of such cancers irrespective of the tissue where they first appear. In order to test if the uncovered GRN core module not only underlies and restricts the types of cell phenotypes (attractors) but also their time–ordered transitions, following [25] we explored its associated Epigenetic Landscape (EL) by implementing a discrete stochastic model as an extension to the Boolean network model [12] (see Methods). By means of computer–based simulations we performed two analyses in order to uncover functional and structural constraints in attractor transitions. (1) We explored the temporal sequence of attractor attainment, and (2) we calculated the consistent global ordering of all the given attractors. Specifically, following [24], we found that the most probable temporal order of attractor attainment for a cell (population) initially on epithelial state is:

$$\text{Epithelial} \rightarrow \text{Senescent} \rightarrow \text{Mesenchymal stem–like},$$

see Fig 4a. On the other hand, following [64] we defined a consistent global ordering of the uncovered attractors based on their relative stability (see Methods). Relative stability calculations are based on the mean first passage time (MFPT) between pairs of attractors. These, in turn, epitomize barrier heights in the EL by approximating a measure for the ease of specific transitions. Similar to the previous analysis, the uncovered global ordering of attractors is Epithelial → Senescent → Mesenchymal stem–like (Fig 4b). This corresponds to the only order in which the system can visit the three attractors following a positive net transition rate. These results indicate that, when considering only intracellular regulatory constraints alone, the uncovered GRN core module structures the epigenetic landscape in a way that a specific flow across the landscape is preferentially and robustly followed. We anticipate that observed transition rates in vivo are likely to depend on tissue–level processes and/or additional GRN components underlying epithelial cell sub–differentiation, that have not been considered here. These latter restrictions will be modeled in future contributions building up on the framework that has been put forward here.

Discussion

Multicellularity by definition implies a one–to–many genotype–phenotype map. The genome of a multicellular individual possesses the intrinsic potentiality to implement a developmental process by which all its different cell–types and tissue structures are ultimately established. In the last decades, a quite coherent theory to explain the development of multicellular organisms as the result of the orchestrating role of GRNs has been developed [9, 11, 12]. The main conclusion is that observable cell states emerge from the self–consistent multistable regulatory logic dictated by genome structure and obeyed by (mainly) transcription factors (TFs) resulting in stable, steady–states of gene expression. Cancer development and progression is also a phenomenon intrinsic to multicellular organisms. Furthermore, similar to normal development, cancer is robustly established as evidenced by its directionality and phenotypic convergence [2]. Is cancer somehow orchestrated by GRN dynamics as well? Several hypothesis have been presented in this direction such as the cancer attractor theory [2,8], and the endogenous molecular cellular network hypothesis [65,66]. In this contribution we also follow the viewpoint of an intrinsic regulatory network, but we focus on a specific developmental process at the cellular level: the robust cell state transitions observed during the tumorigenic transformation of human epithelial cells in culture induced by inflammation and resulting from surpassing a senescent state through EMT – *i.e.,* tumorigenic transformation due to spontaneous immortalization. We propose that a mechanistic understanding of this process is an important first necessary step to unravel key cellular processes which might be occurring *in vivo*, where its rate of occurrence is likely to be regulated by tissue–level and systemic conditions directly linked with lifestyle choices, as well as additional regulatory interactions underlying epithelial cell sub–differentiation.

A Generic Molecular Regulatory Network

The predominant strategy in the molecular study of cancer and cellular tumorigenic transformation has been to focus on pathways and associated mutations. Aware that signaling pathways are actually embedded in complex regulatory networks here we assembled from curated literature a GRN comprising the main molecular regulators involved in key cellular processes ubiquitous to carcinogenesis following a bottom–up approach (see results). Subsequently, we followed a mechanistic approach to address the question of whether we assembled a set of necessary and sufficient molecular players and interactions to recover the cellular phenotypes and processes documented during the spontaneous immortalization of human epithelial cells in culture: we proposed, analyzed and validated an experimentally grounded core GRN dynamical model.

Small developmental regulatory modules have been shown to successfully include the necessary and sufficient set of components and interactions for explaining, as manifestations of intrinsic structural and functional constraints imposed by these GRNs, the dynamics of complex processes such as stem cell differentiation [67], cell–fate decision [68] and similar cellular processes during plant morphogenesis [9,10,22,24]. We hypothesized that a similar core developmental module can be formulated in an attempt to explain the cell–fates observed during spontaneous immortalization of human epithelial cells *in vitro* resulting in a potentially tumorigenic state. In order to show this, we first reduced the proposed larger network into a regulatory core module, by eliminating transitory pathways within the network and by including compounded nodes while maintaining the core network structure and without affecting the dynamical output during each reduction step (for details, see Methods). We obtained a small set of main molecular players (Fig 2). We extracted from available literature the expression profiles of the generally observable cell states of interest in terms of this minimal set of molecules (see Table 2). Given our main hypothesis, we tested if the reduced molecular set and their regulatory logic formalized as a Boolean GRN model were able to recover the biologically observable expression profiles as stationary and stable network configurations (*i.e.*, attractors). Interestingly, we found that the core GRN model only converges to the observed gene expression profiles in wild–type (see Table 2) and some mutant backgrounds (see results).

This result strongly suggest that we have successfully included the key regulators and interactions at play during the establishment of cell states observed during the tumorigenic transformation of human epithelial cells resulting from spontaneous immortalization.

It is noteworthy that our model does not include any hypothetical interaction or component, a common practice in GRN modeling [10, 22, 68]. Our GRN model exclusively integrates available published experimental data; indeed, it was a surprising result that the observed dynamical behavior emerged naturally under such conditions. This suggests that despite incomplete information, there is enough molecular data to uncover important restrictions underlying cell behavior during transitions relevant to epithelial carcinogenesis. Consequently, we consider that the networks reported herein (both the large and the core GRNs) may serve as *bona fide* base models useful to integrate novel discoveries, as well as components underlying epithelial cellular sub–differentiation, while following a bottom–up approach in cancer network systems biology.

Attractor Time–Ordered Transitions

Discrete GRN models can be used to integrate regulatory mechanisms that not only recapitulate the observed gene expression patterns, but that also reproduce the observed developmental time–ordering of cell phenotypes. This can be done by considering stochasticity in the model in order to explore [12,23,25] and/or characterize [64] the associated EL. Importantly, by exploring noise–induced transitions we do not assume that noise alone is the driving force of the transitions, instead, we exploit noise as a tool to explore the GRN–based version of Waddington's EL and to indirectly characterize its structure. Specifically, by calculating the relative stability of the attractors (see Methods) we approximate the in–between attractor barrier heights in the landscape. Furthermore, measures of relative stability can also be exploited to calculate net transition rates measuring the ease of specific inter–attractor transitions and to uncover the predominant developmental route across the epigenetic landscape [69]: ordered transitions sharing positive net transition rates will be preferentially followed. Our results show that such a developmental route follows the time–order of cellular phenotypic states epithelial→senescent→mesenchymal stem–like (potentially tumorigenic). In other words, the constraints imposed by the GRN structure the associated EL in such a way that an epithelial cell in culture as a "ball" would naturally roll following such a path, in agreement with the observed spontaneous immortalization process.

Even in the case of the simple model presented here, it is interesting that of the many possible cell states and developmental routes, the core GRN network is canalized to the few steady–states and the developmental time–ordering consistent with the molecular characterization of cell phenotypes observed during spontaneous immortalization and correlating with carcinoma progression *in vivo* (see below). This suggests that specific progressive alterations or particular "abnormal" signaling mechanisms are not necessarily required for a cell to reach a potentially tumorigenic state. Additionally, robustness analysis performed on the same network showed that the recovered attractors are also robust to permanent alterations of the regulatory logic.

From Abstract Network Attractors and Dynamics to Biological Insight

We are aware of the high degree of simplification involved in the model proposed herein. Accordingly, we do not attempt to present it as a source of accurate predictions for either the occurrence or the future behavior of a phenomena as complex as carcinogenesis. Instead, we formulate the model in an attempt to provide some intuition into otherwise highly complicated processes, and to illuminate increasing body of confounding descriptions. Simple mechanistic models like the one presented here sacrifice detail and accuracy in exchange for understanding [47, 70]. What biological insights can be gained by the uncovered GRN dynamical model? Our simple GRN model strongly suggests that the generic series of cell state transitions widely observed and robustly induced by inflammation in cell culture from normal epithelial to immortalized senescent cells, and from this latter state to a final mesenchymal stem–like phenotype

in the process defined as spontaneous immortalization naturally result from the self–organized behavior emerging from an underlying GRN novel architecture and topology.

Importantly, cells that emerge from spontaneous immortalization induced by cytokines display mesenchymal stem like phenotype and tumorigenic behavior – *i.e.*, repress proteins p16 and p53, surpass senescence, and re–express telomerase [18]. Phenotypically, these cells are difficult to distinguish from the so–called cancer stem cells, tumor initiating cells or embryonic stem cells [20, 21]; are resistant to apoptosis; and have the ability to migrate and generate metastasis and form secondary tumors – all lethal traits characterizing cancer cells [3]. We, thus, speculate that tissue–level conditions associated with a bad prognosis, such as a pro–inflammatory milieu, may increase the rate of occurrence of these same transitions *in vivo* promoting as a result the development and progression of epithelial cancer. We substantiate this view by noting several independent empirical observations. (1) Histological diagnosis of carcinoma are generally preceded by a lesion called hyperplasia; senescent cells are abundant in hyperplasias and scarce in carcinomas [71]. (2) During chronological aging senescent cells increase in number within both normal tissues and hyperplasias. (3) Senescence is associated with the promotion of carcinogenesis by contributing with the loss of tissue architecture and promoting an inflammatory milieu [72]. (3) Overcoming the senescent barrier is fundamental in tumor progression [73, 74]. (4) The EMT process constitutes a well–characterized mean to overcome senescence under an inflammatory environment([75]).

We must point out, however, that transition rates during spontaneous immortalization, if occurring *in vivo*, may be regulated by tissue–level, self–organizational processes not considered in our cellular level model. For example, the likelihood of spontaneous immortalization *in vivo* may be increased by extracellular perturbations that inevitably occur during aging; mainly, by inflammation and tissue remodeling resulting from an increased population of senescent cells. The cellular level network models reported here are, nevertheless, a valuable building block for more detailed multi–level models integrating further sources of tissue–level constraints such as cell cycle progression, cell–cell interactions, differential proliferation rates, and mechanical forces.

Summarizing, in this contribution we propose an experimentally grounded GRN model for spontaneous immortalization. We report one large GRN model (41 nodes) and one core GRN developmental module (9 nodes), both useful and necessary for further integration of signaling and mechanical processes in multi–level, more detailed modeling efforts. We explore by analyzing the dynamical behavior of the latter if the uncovered GRN topology and architecture underlies the gene expression configurations that characterize normal epithelial, senescent, and mesenchymal stem–like cell–fates well documented during tumorigenic transformation *in vitro* and which correlate with those observed in the progressive pathological description of epithelial carcinogenesis *in vivo*. Overall, our results suggest that tumorigenic transformation *in vitro* due to spontaneous immortalization can be understood and modeled at a cellular level generically as a developmental system undergoing cell–state transitions resulting from the structural and functional constraints imposed, in part, by the interactions included in the proposed GRN. They also suggest that similar transitions may be occurring *in vivo* and might be relevant for carcinoma development and progression. This view is consistent with the robustness, generic patterns, and directionality observed during the development of human cancers derived from epithelial tissues. Particularly, based on our results, we hypothesize that replicative senescence and chronic inflammation are likely to increase the occurrence of spontaneous immortalization *in vivo* promoting the development of epithelial carcinogenesis. Testing such hypothesis awaits the development of multi–level models taking the ones presented here as building blocks, and is the subject of ongoing investigation.

Materials and Methods

Literature Search

A total of 159 references, considering both references in extended view material (see Text S1) and main text, were carefully and manually reviewed in order to first define a minimal set of cellular phenotypes and processes (for definitions, see Text S1) which enable a generic representation of epithelial carcinogenesis on the basis of cell state transition events. Subsequently, a set of associated, experimentally described molecular regulators was extracted from the literature, including their regulatory interactions.

Network Assembly

The network (see Fig. 1) was assembled manually by adding nodes (genes/proteins) and edges (activating or inhibitory interactions) describing direct mechanisms reported in the available literature to have an influence on both the specification of the cellular phenotypes and the development of the cellular process defined in (Text S1). The initial network was created based on experimentally grounded knowledge from 159 references (including reviews and research papers) and consists of 41 nodes and 97 edges. The literature included data known before 2014. Support for each of the proposed interactions is listed in Text S1.

Network–based Gene Set Enrichment Analysis

The bioinformatics tools EnrichNet [76] and TopoGSA [77] were used to perform network–based gene set enrichment analysis and topology–based gene set analysis, respectively. Briefly, EnrichNet maps the input gene set into a molecular interaction network and calculates distances between the genes and pathways/processes in a reference database. TopoGSA also maps the input gene set into a network, and then it computes its topological statistics and compares it against the topology of pathways/processes in a reference database. Here a connected human interactome graph extracted from the STRING database and the KEGG and GO Biological Process databases were used as reference molecular interaction network and databases. Both analyses were performed using the Cytoscape plugin Jepettp [78].

Network Reduction

In order to extract a representative core regulatory model from the initial network and to obtain a more computationally tractable one, which reasonably unfolds the regulatory pathways, a reduction methodology was followed based on certain simplifying assumptions – supported by previous results in molecular biology studies – and on mathematical results from dynamical systems and graph theory. Here we briefly describe the main steps. The step–by–step reduction process is included in Text S1.

Simplifying assumptions:

- ESE-2 groups activities of ESE-1, ESE-3, EGF, Her-2/neu.
- Snai2 groups activities of Snail, Twist (Twist, in turn, groups activities of Twist1 and Twist2), Zeb and FOXC2.
- p16 groups p14 and NF–κB node groups the inflammatory response activated by growth factors, mitogens and cytokines.

Reduction process: (1) Simple mediator nodes (*i.e.*, those nodes with in–degree and out–degree of one) were removed iteratively. (2) Nodes with in–degree of one and out–degree greater than one were removed iteratively. These steps (1 and 2) does not alter the attractors of the Boolean network under the asynchronous update, as mathematically proved in [79]. (3) Redundant interactions of selected nodes (based on biological arguments) resulting in self–regulation were included in single nodes/interactions (for details, see Text S1). (4) Selected nodes (based on biological knowledge again) with in–degree greater than one and out–degree of one were removed. The final steps (3 and 4) are supported by the mathematical analysis made in [80] in which the authors prove that the methodology preserves relevant topological and dynamical properties.

It is noteworthy that fixed point attractors are time–independent, so they are the same in both synchronous and asynchronous update methods. Complex attractors (in which the system oscillates among a set of states), on the other hand, depend on the update method. Consistently, the update method used in the model is irrelevant for the obtained results. This last assertion is valid because the model shows only fixed point attractors, which means, under the mathematically proved reduction methods applied, that *the large network describes a qualitative long time behavior conserved in the reduced one*. Besides, the methodology applied in order to obtain the reduced network enables the analysis of a resulting regulatory graph which is biologically meaningful and dynamically consistent with the network constructed with available molecular biology experimental data.

The final reduced network is shown in Figure 2. We refer to this network and its corresponding logical rules as the core regulatory module.

Dynamical Gene Regulatory Network Model

A Boolean network models a dynamical system assuming both discrete time and discrete state variables. This is expressed formally with the mapping:

$$x_i(t+1) = F_i(x_1(t), x_2(t), ..., x_k(t)), \tag{1}$$

where the set of functions F_i are logical prepositions (or truth tables) expressing the relationship between the genes that share regulatory interactions with the gene i, and where the state variables $x_i(t)$ can take the discrete values 1 or 0 indicating whether the gene i is expressed or not at a certain time t, respectively. An experimentally grounded Boolean GRN model is then completely specified by the set of genes proposed to be involved in the process of interest and the associated set of logical functions derived from experimental data [23]. The set of logical functions for the core regulatory module used in this study is included in Text S1 – both as logical prepositions and truth tables. The dynamical analysis of the Boolean network model was conducted using the package *BoolNet* [63] within the *R* statistical programming environment (www.R-project.org).

Epigenetic Landscape Exploration

Including Stochasticity

In order to extend the Boolean Network into a discrete stochastic model and then study the properties of its associated EL, the so–called stochasticity in nodes (SIN) model was implemented following [23–25]. In this model, a constant probability of error ξ is introduced for the deterministic Boolean functions. In other words, at each time step, each gene "disobeys" its Boolean function with probability ξ. Formally:

$$\begin{aligned} P_{x_i(t+1)}[F_i(\mathbf{x}_{reg_i}(t))] &= 1-\xi, \\ P_{x_i(t+1)}[1-F_i(\mathbf{x}_{reg_i}(t))] &= \xi. \end{aligned} \tag{2}$$

The probability that the value of the now random variable $x_i(t+1)$ is determined or not by its associated logical function $F_i(\mathbf{x}_{reg_i}(t))$ is $1-\xi$ or ξ, respectively.

Attractor Transition Probability Estimation

An attractor transition probability matrix Π with components:

$$\pi_{ij} = P(A_{t+1} = j | A_t = i), \tag{3}$$

representing the probability that an attractor j is reached from an attractor i, was estimated by numerical simulation following [24]. Specifically, for each network state i in the state space (2^n) a stochastic one–step transition was simulated a large number of times ($\approx 10,000$). The probability of transition from an attractor i to an attractor j was then estimated as the frequency of times the states belonging to the basin of the attractor i were stochastically mapped into a state within the basin of the attractor j.

Following the discrete time Markov chains (DTMCs) [81] theoretical framework, the estimated transition probability matrix was integrated into a dynamic equation for the probability distribution:

$$P_A(t+1) = \Pi P_A(t), \tag{4}$$

where $P_A(t)$ is the probability distribution over the attractors at time t, and Π is the transition probability matrix. This equation was iterated to simulate the temporal evolution of the probability distribution over the attractors starting from a specific initial probability distribution.

Attractor Relative Stability and Global Ordering Analyses

In addition to the calculation of the most probable temporal cell–fate pattern (see [24]), a discrete stochastic GRN model enables the study of the ease for transitioning from one attractor to another [69]. Specifically, a transition barrier in the EL epitomizes the ease for transitioning from one attractor to another. The ease of transitions, in turn, offers a notion of relative stability. It has recently been proposed that the GRN has a consistent global ordering of all cell attractors and intermediate transient states which can be uncovered by measuring the relative stabilities of all the attractors of a Boolean GRN [64, 69]. Here, the relative stabilities of the cell states were defined based on the mean first passage time (MFPT). Specifically, a relative stability matrix M was calculated which reflects the transition barrier between any two states based on the MFPT. Here, in all cases, the MFPT was estimated numerically. Using the transition probabilities among attractors, a large number sample paths of a finite Markov chain were simulated. The MFPT from attractor i to attractor j corresponds to the averaged value of the number of steps taken to visit attractor j for the fist time, given that the entire probability mass was initially localized at the attractor i. The average is taken over the realizations. Following [69], based on the MFPT values a net transition rate between attractor i and j can be defined as follows:

$$d_{i,j} = \frac{1}{MFPT_{i,j}} - \frac{1}{MFPT_{j,i}} \tag{5}$$

This quantity effectively measures the ease of transition as a net probability flow. For all the calculation involving stochasticity, the robustness of the results was assessed by taking three different values for the probability of error $(0.01, 0.05, 0.1)$. Stability of the results was assessed by manually changing the number of simulated samples until results become stable.

The consistent global ordering of all attractors uncovered with the core GRN was defined based on the formula proposed in [64]. Briefly, the consistent global ordering of the attractors is given by the attractor permutation in which al transitory net transition rates from an initial attractor to a final attractor are positive. This is schematically represented in Figure 4b. Calculated transition probability, MFPT, and net transition rate matrices are included in Text S2. R source code implementing all the calculations and analyses is available upon request.

Authors' contributions

ERAB and JMG coordinated the study and with the other authors established the overall logic and core questions to be addressed. All the authors conceived and planned the modeling approaches. FML recovered the information from the literature to establish the model and provided expert knowledge in cancer biology. JDV established many of the specific analyses to be done, and programmed and ran all the modeling and analyses. FML and CEO formalized experimental data into regulatory logic. ERAB, JMG, JDV and FML participated in the interpretation of the results and analyses. JDV wrote most of the paper with help from ERAB and JMG and input from FML. All authors proofread the final version of the ms submitted.

Acknowledgments

This work was supported by grants CONACYT 240180, 180380, 167705, 152649 and UNAM-DGAPA-PAPIIT: IN203113, IN 203214, IN203814, UC Mexus ECO-IE415. J.D.V acknowledges the support of CONACYT and the Centre for Genomic Regulation (CRG), Barcelona, Spain; while spending a research visit in the lab of Stephan Ossowski. This article constitutes a partial fulfillment of the graduate program Doctorado en Ciencias Biomédicas of the Universidad Nacional Autónoma de México, UNAM in which J.D.V. developed this project. J.D.V receives a PhD scholarship from CONACYT. The authors acknowledge logistical and administrative help of Diana Romo.

References

1. Anand P, Kunnumakara AB, Sundaram C, Harikumar KB, Tharakan ST, Lai OS, et al. Cancer is a preventable disease that requires major lifestyle changes. Pharmaceutical research. 2008;25(9):2097–2116.

2. Huang S. On the intrinsic inevitability of cancer: from foetal to fatal attraction. In: Seminars in cancer biology. vol. 21. Elsevier; 2011. p. 183–199.

3. Mani SA, Guo W, Liao MJ, Eaton EN, Ayyanan A, Zhou AY, et al. The epithelial-mesenchymal transition generates cells with properties of stem cells. Cell. 2008;133(4):704–715.

4. Huang S. Non-genetic heterogeneity of cells in development: more than just noise. Development. 2009;136(23):3853–3862.

5. Ben-Porath I, Thomson MW, Carey VJ, Ge R, Bell GW, Regev A, et al. An embryonic stem cell–like gene expression signature in poorly differentiated aggressive human tumors. Nature genetics. 2008;40(5):499–507.

6. Kelloff GJ, Sigman CC. Assessing intraepithelial neoplasia and drug safety in cancer-preventive drug development. Nature Reviews Cancer. 2007;7(7):508–518.

7. Virchow RLK. Cellular pathology. John Churchill; 1860.

8. Huang S, Ernberg I, Kauffman S. Cancer attractors: a systems view of tumors from a gene network dynamics and developmental perspective. In: Seminars in cell & developmental biology. vol. 20. Elsevier; 2009. p. 869–876.

9. Mendoza L, Alvarez-Buylla ER. Dynamics of the genetic regulatory network for Arabidopsis thaliana flower morphogenesis. J Theor Biol. 1998;193(2):307–319.

10. Espinosa-Soto C, Padilla-Longoria P, Alvarez-Buylla ER. A gene regulatory network model for cell-fate determination during Arabidopsis thaliana flower development that is robust and recovers experimental gene expression profiles. The Plant Cell Online. 2004;16(11):2923–2939.

11. Huang S, Kauffman S. Complex gene regulatory networks-from structure to biological observables: cell fate determination. Encyclopedia of Complexity and Systems Science Meyers RA, editors Springer. 2009;p. 1180–1293.

12. Alvarez-Buylla ER, Azpeitia E, Barrio R, Benítez M, Padilla-Longoria P. From ABC genes to regulatory networks, epigenetic landscapes and flower morphogenesis: making biological sense of theoretical approaches. Seminars in cell & developmental biology. 2010;21(1):108–117.

13. Huang S. Reprogramming cell fates: reconciling rarity with robustness. Bioessays. 2009;31(5):546–560.

14. Kaneko K. Characterization of stem cells and cancer cells on the basis of gene expression profile stability, plasticity, and robustness. Bioessays. 2011;33(6):403–413.

15. Huang S. The molecular and mathematical basis of Waddington's epigenetic landscape: A framework for post-Darwinian biology? Bioessays. 2012;34(2):149–157.

16. Davila-Velderrain J, Alvarez-Buylla ER. Bridging the Genotype and the Phenotype: Towards An Epigenetic Landscape Approach to Evolutionary Systems Biology. bioRxiv. 2014;.

17. Xu J, Lamouille S, Derynck R. TGF-β-induced epithelial to mesenchymal transition. Cell research. 2009;19(2):156–172.

18. Battula VL, Evans KW, Hollier BG, Shi Y, Marini FC, Ayyanan A, et al. Epithelial-Mesenchymal Transition-Derived Cells Exhibit Multilineage Differentiation Potential Similar to Mesenchymal Stem Cells. Stem Cells. 2010;28(8):1435–1445.

19. Li CW, Xia W, Huo L, Lim SO, Wu Y, Hsu JL, et al. Epithelial–mesenchymal transition induced by TNF-α requires NF-κB–mediated transcriptional upregulation of Twist1. Cancer research. 2012;72(5):1290–1300.

20. Morel AP, Lièvre M, Thomas C, Hinkal G, Ansieau S, Puisieux A. Generation of breast cancer stem cells through epithelial-mesenchymal transition. PloS one. 2008;3(8):e2888.

21. Neph S, Stergachis AB, Reynolds A, Sandstrom R, Borenstein E, Stamatoyannopoulos JA. Circuitry and dynamics of human transcription factor regulatory networks. Cell. 2012;150(6):1274–1286.

22. Azpeitia E, Benítez M, Vega I, Villarreal C, Alvarez-Buylla ER. Single-cell and coupled GRN models of cell patterning in the Arabidopsis thaliana root stem cell niche. BMC systems biology. 2010;4(1):134.

23. Azpeitia E, Davila-Velderrain J, Villarreal C, Alvarez-Buylla ER. Gene regulatory network models for floral organ determination. In: Flower Development. Springer; 2014. p. 441–469.

24. Álvarez-Buylla ER, Chaos Á, Aldana M, Benítez M, Cortes-Poza Y, Espinosa-Soto C, et al. Floral morphogenesis: stochastic explorations of a gene network epigenetic landscape. Plos one. 2008;3(11):e3626.

25. Davila-Velderrain J, Martínez-García J, Alvarez-Buylla ER. Modeling the Epigenetic Attractors Landscape: Towards a Post-Genomic Mechanistic Understanding of Development. Name: Frontiers in Genetics. 2015;6:160.

26. Zhou J, Ng A, Tymms MJ, Jermiin LS, Seth AK, Thomas RS, et al. A novel transcription factor, ELF5, belongs to the ELF subfamily of ETS genes and maps to human chromosome 11p13-15, a region subject to LOH and rearrangement in human carcinoma cell lines. Oncogene. 1998;17(21):2719–2732.

27. Ma XJ, Salunga R, Tuggle JT, Gaudet J, Enright E, McQuary P, et al. Gene expression profiles of human breast cancer progression. Proceedings of the National Academy of Sciences. 2003;100(10):5974–5979.

28. Escamilla-Hernandez R, Chakrabarti R, Romano RA, Smalley K, Zhu Q, Lai W, et al. Genome-wide search identifies Ccnd2 as a direct transcriptional target of Elf5 in mouse mammary gland. BMC molecular biology. 2010;11(1):68.

29. Chakrabarti R, Hwang J, Blanco MA, Wei Y, Lukačišin M, Romano RA, et al. Elf5 inhibits the epithelial–mesenchymal transition in mammary gland development and breast cancer metastasis by transcriptionally repressing Snail2. Nature cell biology. 2012;14(11):1212–1222.

30. Vernier M, Bourdeau V, Gaumont-Leclerc MF, Moiseeva O, Bégin V, Saad F, et al. Regulation of E2Fs and senescence by PML nuclear bodies. Genes & development. 2011;25(1):41–50.

31. McConnell BB, Gregory FJ, Stott FJ, Hara E, Peters G. Induced expression of p16 INK4a inhibits both CDK4-and CDK2-associated kinase activity by reassortment of cyclin-CDK-inhibitor complexes. Molecular and cellular biology. 1999;19(3):1981–1989.

32. Villacañas Ó, Pérez JJ, Rubio-Martínez J. Structural analysis of the inhibition of Cdk4 and Cdk6 by p16INK4a through molecular dynamics simulations. Journal of Biomolecular Structure and Dynamics. 2002;20(3):347–358.

33. Bracken AP, Kleine-Kohlbrecher D, Dietrich N, Pasini D, Gargiulo G, Beekman C, et al. The Polycomb group proteins bind throughout the INK4A-ARF locus and are disassociated in senescent cells. Genes & development. 2007;21(5):525–530.

34. Fang L, Igarashi M, Leung J, Sugrue MM, Lee SW, Aaronson SA. p21Waf1/Cip1/Sdi1 induces permanent growth arrest with markers of replicative senescence in human tumor cells lacking functional p53. Oncogene. 1999;18(18):2789–2797.

35. Mao Z, Ke Z, Gorbunova V, Seluanov A. Replicatively senescent cells are arrested in G1 and G2 phases. Aging (Albany NY). 2012;4(6):431.

36. Chellappan SP, Hiebert S, Mudryj M, Horowitz JM, Nevins JR. The E2F transcription factor is a cellular target for the RB protein. Cell. 1991;65(6):1053–1061.

37. Byeon IJL, Li J, Ericson K, Selby TL, Tevelev A, Kim HJ, et al. Tumor Suppressor p16INK4A: Determination of Solution Structure and Analyses of Its Interaction with Cyclin-Dependent Kinase 4. Molecular cell. 1998;1(3):421–431.

38. Beauséjour CM, Krtolica A, Galimi F, Narita M, Lowe SW, Yaswen P, et al. Reversal of human cellular senescence: roles of the p53 and p16 pathways. The EMBO journal. 2003;22(16):4212–4222.

39. Freudlsperger C, Bian Y, Wise SC, Burnett J, Coupar J, Yang X, et al. TGF-β and NF-κB signal pathway cross-talk is mediated through TAK1 and SMAD7 in a subset of head and neck cancers. Oncogene. 2012;32(12):1549–1559.

40. Harley C, Futcher A, Greider C. Telomeres shorten during ageing of human fibroblasts. Nature. 1990;345(6274):458–460.

41. Mani SA, Yang J, Brooks M, Schwaninger G, Zhou A, Miura N, et al. Mesenchyme Forkhead 1 (FOXC2) plays a key role in metastasis and is associated with aggressive basal-like breast cancers. Proceedings of the National Academy of Sciences. 2007;104(24):10069–10074.

42. Zeisberg M, Neilson EG, et al. Biomarkers for epithelial-mesenchymal transitions. The Journal of clinical investigation. 2009;119(6):1429–1437.

43. Bolós V, Peinado H, Pérez-Moreno MA, Fraga MF, Esteller M, Cano A. The transcription factor Slug represses E-cadherin expression and induces epithelial to mesenchymal transitions: a comparison with Snail and E47 repressors. Journal of cell science. 2003;116(3):499–511.

44. Dave N, Guaita-Esteruelas S, Gutarra S, Frias À, Beltran M, Peiró S, et al. Functional cooperation between Snail1 and twist in the regulation of ZEB1 expression during epithelial to mesenchymal transition. Journal of Biological Chemistry. 2011;286(14):12024–12032.

45. Casas E, Kim J, Bendesky A, Ohno-Machado L, Wolfe CJ, Yang J. Snail2 is an essential mediator of Twist1-induced epithelial mesenchymal transition and metastasis. Cancer research. 2011;71(1):245–254.

46. Hajra KM, Chen DY, Fearon ER. The SLUG zinc-finger protein represses E-cadherin in breast cancer. Cancer research. 2002;62(6):1613–1618.

47. Davila-Velderrain J, Martinez-Garcia J, Alvarez-Buylla E. Descriptive vs. Mechanistic Network Models in Plant Development in the Post-Genomic Era. Plant Functional Genomics: Methods and Protocols. 2015;p. 455–479.

48. Siegel R, Naishadham D, Jemal A. Cancer statistics, 2013. CA: a cancer journal for clinicians. 2013;63(1):11–30.

49. Jemal A, Bray F, Center MM, Ferlay J, Ward E, Forman D. Global cancer statistics. CA: a cancer journal for clinicians. 2011;61(2):69–90.

50. Stratton MR, Campbell PJ, Futreal PA. The cancer genome. Nature. 2009;458(7239):719–724.

51. Hudson TJ, Anderson W, Aretz A, Barker AD, Bell C, Bernabé RR, et al. International network of cancer genome projects. Nature. 2010;464(7291):993–998.

52. Weinstein JN, Collisson EA, Mills GB, Shaw KRM, Ozenberger BA, Ellrott K, et al. The cancer genome atlas pan-cancer analysis project. Nature genetics. 2013;45(10):1113–1120.

53. Yaffe MB. The scientific drunk and the lamppost: massive sequencing efforts in cancer discovery and treatment. Science signaling. 2013;6(269):pe13.

54. Creixell P, Schoof EM, Erler JT, Linding R. Navigating cancer network attractors for tumor-specific therapy. Nature biotechnology. 2012;30(9):842–848.

55. DePinho RA, Polyak K. Cancer chromosomes in crisis. Nature genetics. 2004;36(9):932–934.

56. Fujikawa M, Katagiri T, Tugores A, Nakamura Y, Ishikawa F. ESE-3, an Ets family transcription factor, is up-regulated in cellular senescence. Cancer science. 2007;98(9):1468–1475.

57. Lee JM, Dedhar S, Kalluri R, Thompson EW. The epithelial–mesenchymal transition: new insights in signaling, development, and disease. The Journal of cell biology. 2006;172(7):973–981.

58. Cano A, Pérez-Moreno MA, Rodrigo I, Locascio A, Blanco MJ, del Barrio MG, et al. The transcription factor snail controls epithelial–mesenchymal transitions by repressing E-cadherin expression. Nature cell biology. 2000;2(2):76–83.

59. Sun Y, Song GD, Sun N, Chen JQ, Yang SS. Slug overexpression induces stemness and promotes hepatocellular carcinoma cell invasion and metastasis. Oncology Letters. 2014;7(6):1936–1940.

60. Liu Y, El-Naggar S, Darling DS, Higashi Y, Dean DC. Zeb1 links epithelial-mesenchymal transition and cellular senescence. Development. 2008;135(3):579–588.

61. Weinberg RA. Twisted epithelial–mesenchymal transition blocks senescence. Nature cell biology. 2008;10(9):1021–1023.

62. Kim WY, Sharpless NE. The Regulation of¡ i¿ INK4¡/i¿/¡ i¿ ARF¡/i¿ in Cancer and Aging. Cell. 2006;127(2):265–275.

63. Müssel C, Hopfensitz M, Kestler HA. BoolNet—an R package for generation, reconstruction and analysis of Boolean networks. Bioinformatics. 2010;26(10):1378–1380.

64. Zhou JX, Samal A, d'Hèrouël AF, Price ND, Huang S. Relative Stability of Network States in Boolean Network Models of Gene Regulation in Development. arXiv preprint arXiv:14076117. 2014;.

65. Wang G, Zhu X, Gu J, Ao P. Quantitative implementation of the endogenous molecular–cellular network hypothesis in hepatocellular carcinoma. Interface focus. 2014;4(3):20130064.

66. Zhu X, Yuan R, Hood L, Ao P. Endogenous molecular-cellular hierarchical modeling of prostate carcinogenesis uncovers robust structure. Progress in biophysics and molecular biology. 2015;.

67. Li C, Wang J. Quantifying cell fate decisions for differentiation and reprogramming of a human stem cell network: landscape and biological paths. PLoS computational biology. 2013;9(8):e1003165.

68. Zhou JX, Brusch L, Huang S. Predicting pancreas cell fate decisions and reprogramming with a hierarchical multi-attractor model. PloS one. 2011;6(3):e14752.

69. Zhou JX, Qiu X, d'Herouel AF, Huang S. Discrete Gene Network Models for Understanding Multicellularity and Cell Reprogramming: From Network Structure to Attractor Landscapes Landscape. In: Computational Systems Biology Second Edition Elsevier. 2014;p. 241–276.

70. Lander AD. The edges of understanding. BMC biology. 2010;8(1):40.

71. Chen Z, Trotman LC, Shaffer D, Lin HK, Dotan ZA, Niki M, et al. Crucial role of p53-dependent cellular senescence in suppression of Pten-deficient tumorigenesis. Nature. 2005;436(7051):725–730.

72. Campisi J. Cellular senescence: putting the paradoxes in perspective. Current opinion in genetics & development. 2011;21(1):107–112.

73. Narita M, Lowe SW. Senescence comes of age. Nature medicine. 2005;11(9):920–922.

74. Yildiz G, Arslan-Ergul A, Bagislar S, Konu O, Yuzugullu H, Gursoy-Yuzugullu O, et al. Genome-wide transcriptional reorganization associated with senescence-to-immortality switch during human hepatocellular carcinogenesis. PloS one. 2013;8(5):e64016.

75. Smit MA, Peeper DS. Epithelial-mesenchymal transition and senescence: two cancer-related processes are crossing paths. Aging (Albany NY). 2010;2(10):735.

76. Glaab E, Baudot A, Krasnogor N, Schneider R, Valencia A. EnrichNet: network-based gene set enrichment analysis. Bioinformatics. 2012;28(18):i451–i457.

77. Glaab E, Baudot A, Krasnogor N, Valencia A. TopoGSA: network topological gene set analysis. Bioinformatics. 2010;26(9):1271–1272.

78. Winterhalter C, Widera P, Krasnogor N. JEPETTO: a Cytoscape plugin for gene set enrichment and topological analysis based on interaction networks. Bioinformatics. 2014;30(7):1029–1030.

79. Saadatpour A, Albert R, Reluga TC. A reduction method for Boolean network models proven to conserve attractors. SIAM Journal on Applied Dynamical Systems. 2013;12(4):1997–2011.

80. Naldi A, Remy E, Thieffry D, Chaouiya C. Dynamically consistent reduction of logical regulatory graphs. Theoretical Computer Science. 2011;412(21):2207–2218.

81. Allen LJ. An introduction to stochastic processes with applications to biology. CRC Press; 2010.

82. Li J, Poi MJ, Tsai MD. Regulatory mechanisms of tumor suppressor P16INK4A and their relevance to cancer. Biochemistry. 2011;50(25):5566–5582.

83. Yamakoshi K, Takahashi A, Hirota F, Nakayama R, Ishimaru N, Kubo Y, et al. Real-time in vivo imaging of p16Ink4a reveals cross talk with p53. The Journal of cell biology. 2009;186(3):393–407.

Figure Legends

Figure 1. Gene regulatory network for epithelial carcinogenesis. Nodes represent genes, and arrows (bars) represent experimentally characterized activation (arrow-heads) or repression (flat-heads) interactions. Genes corresponding to TFs are represented by squares and the rest by circles. (a) Colors indicate association with specific phenotypes and processes: epithelial (green), mesenchymal (orange), inflammation (red), senescence and DNA damage (blue), cell–cycle (purple), and polycomb complex (yellow). (b) Core gene regulatory module in the context of the global network. Colored nodes represent the final set of molecules obtained after the network reduction methodology was applied (see Methods) and which were included in the core GRN model.

Figure 2. Core gene regulatory network module for epithelial carcinogenesis Nodes represent either single or subsets of genes (see Results); arrows-heads represent activations and flat–heads repression interactions. Five of the nodes are involved in the specification of the cellular phenotypes: Epithelial (Ese–2), Senescent (p16, p53), and Mesenchymal stem–like (Snai2, TELasa). Three nodes are tightly associated with cell–cycle regulation (Rb, E2F, Cyclin), while node NF–κB represents cellular inflammation.

Figure 3. The core gene regulatory module in the context of the *Hallmarks of Cancer* approach. The antagonistic activity state ESE–2 (-) and Snai2 (+) enable cells to *sustain proliferative signals* and *evade growth suppressors* by undergoing a dedifferentiation process. The state p16(-), Rb(-), p53(-), and TELasa (+) enable cell to *acquire replicative immortality*, *resist cell death*, as well as present *genome instability* and a *mutation–prone* phenotype by surpassing cellular senescence. High levels of cytokines and NF–κB(+) expose cells to *tumor promoting inflammation*. The constitutive activity of Snai2(+) epitomizes the intrinsic phenotypic features of the cells emerging from the process of inflammation–induced EMT: *activating invasion*, *avoiding immune destruction*, and *deregulating cellular energetics*.

Figure 4. Temporal sequence and global order of cell–fate attainment pattern under the stochastic Boolean GRN model during epithelial carcinogenesis. (a) Maximum probability p of attaining each attractor, as a function of time (in iteration steps). Vertical lines mark the time when maximal probability of each attractor occurs. The most probable sequence of cell attainment is: epithelial(E) $\rightarrow$ senescent(S) $\rightarrow$ mesenchymal(cancer–like)(M). The value of the error probability used in this case was $\xi = 0.05$. The same patterns were obtained with the 3 different error probabilities tested (data not shown). (b) Schematic representation of the possible transitions between pairs of attractors. Arrows indicate the directionality of the transitions. Above each arrow a sign (+) or (−) indicates whether the calculated net transition rate between the corresponding attractors is positive or negative. Red arrows represent the globally consistent ordering for the 3 attractors: the order of the attractors in which all individual transition has a positive net rate, resulting in a global probability flow across the EL.

Supporting Information captions

Text S1. Supplementary text including detailed methodology and definitions.

Text S2. Supplementary text including calculated transition probability, MFPT, and net transition rate matrices.

Figure S1. Network topological gene set analysis results.

Figure S2. Recovered attractors in mutant conditions.

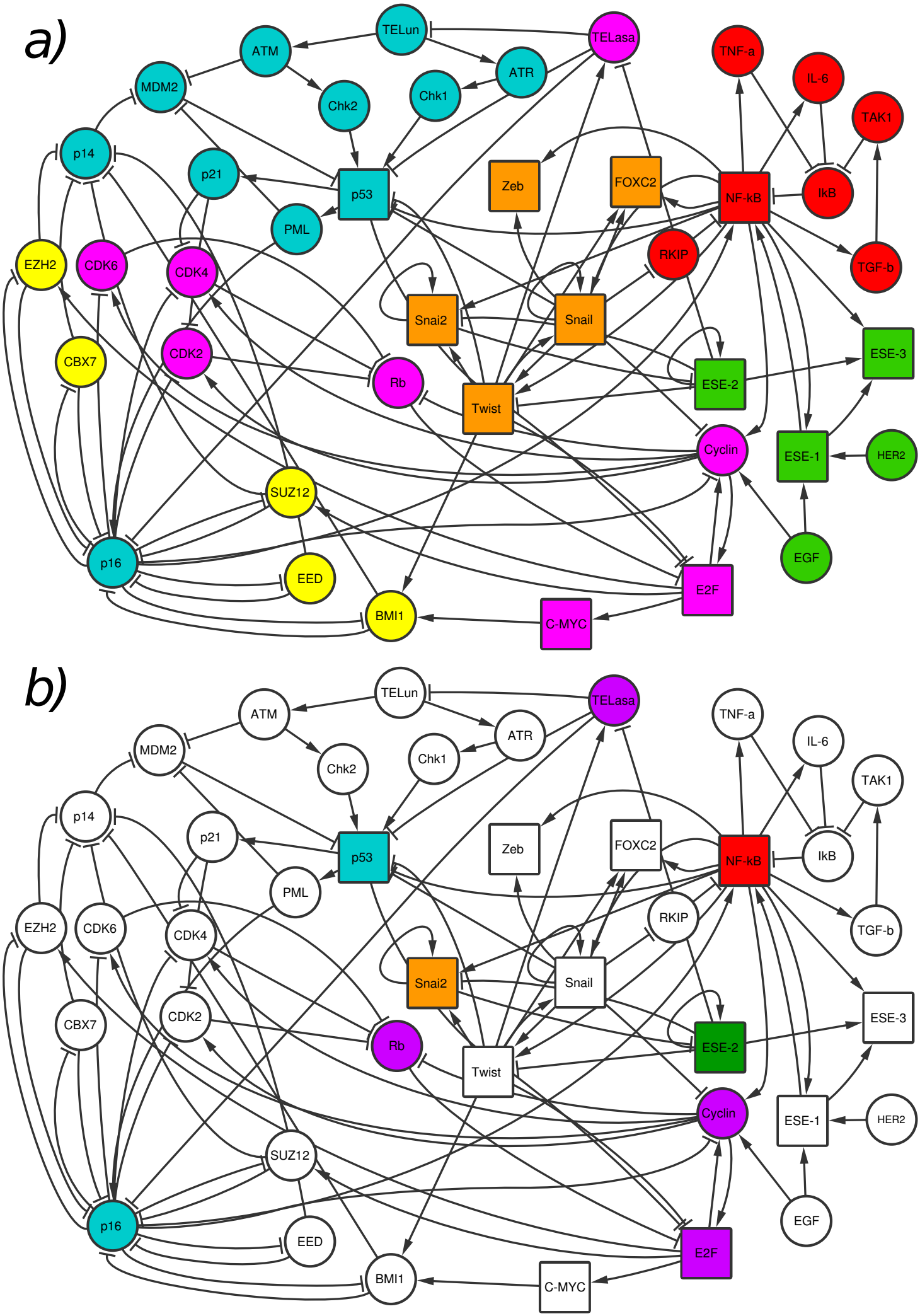
Tables

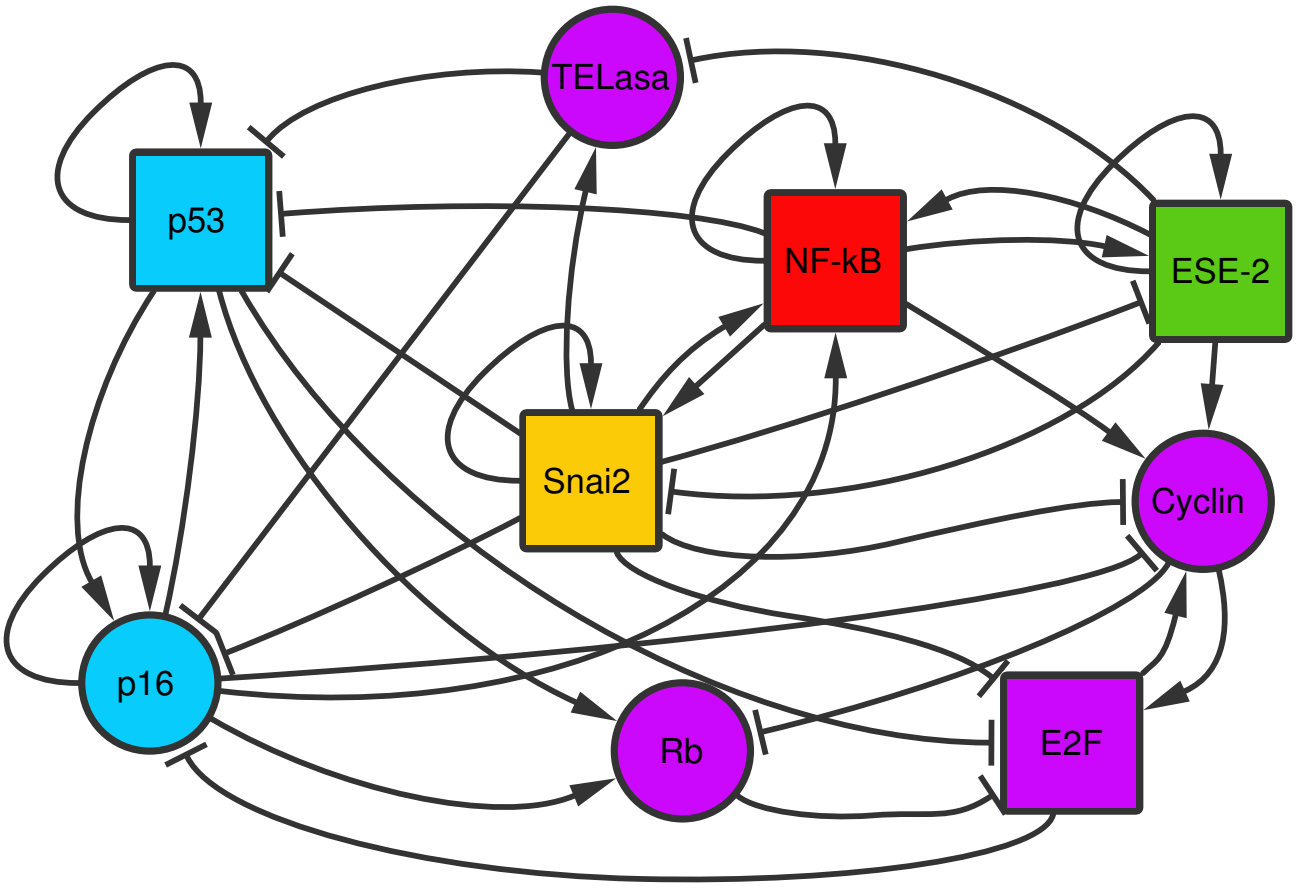
KEGG – Pathway or Process	XD–score	q-value	Overlap/Size
Bladder cancer	1.1447	0	12/38
Chronic myeloid leukemia	0.86866	0	17/69
p53 signaling pathway	0.78477	0	14/62
Pancreatic cancer	0.68155	0	14/70
Glioma	0.68155	0	12/60
Non–small cell lung cancer	0.66586	0	10/51
Melanoma	0.65574	0	12/62
Small cell lung cancer	0.56447	0	14/82
Prostate cancer	0.54821	0	14/84
Cell cycle	0.54821	0	20/120
Cytosolic DNA–sensing pathway	0.48155	0.00001	6/40
Thyroid cancer	0.36155	0.00784	3/25
NOD-like receptor signaling pathway	0.35612	0.00001	7/59
GO Biological Process	**XD-score**	**q–value**	**Overlap/Size**
replicative senescence	3.13328	0	8/10
cellular senescence	0.73328	0.02244	2/10
cell aging	0.43328	0.00608	3/24
activation of NF–κB–inducing kinase activity	0.43328	0.04656	2/16
determination of adult lifespan	0.33328	0.40382	1/10
epithelial cell differentiation	0.32721	0.13188	2/33
positive regulation of NF–κB transcription factor activity	0.30109	0	8/87

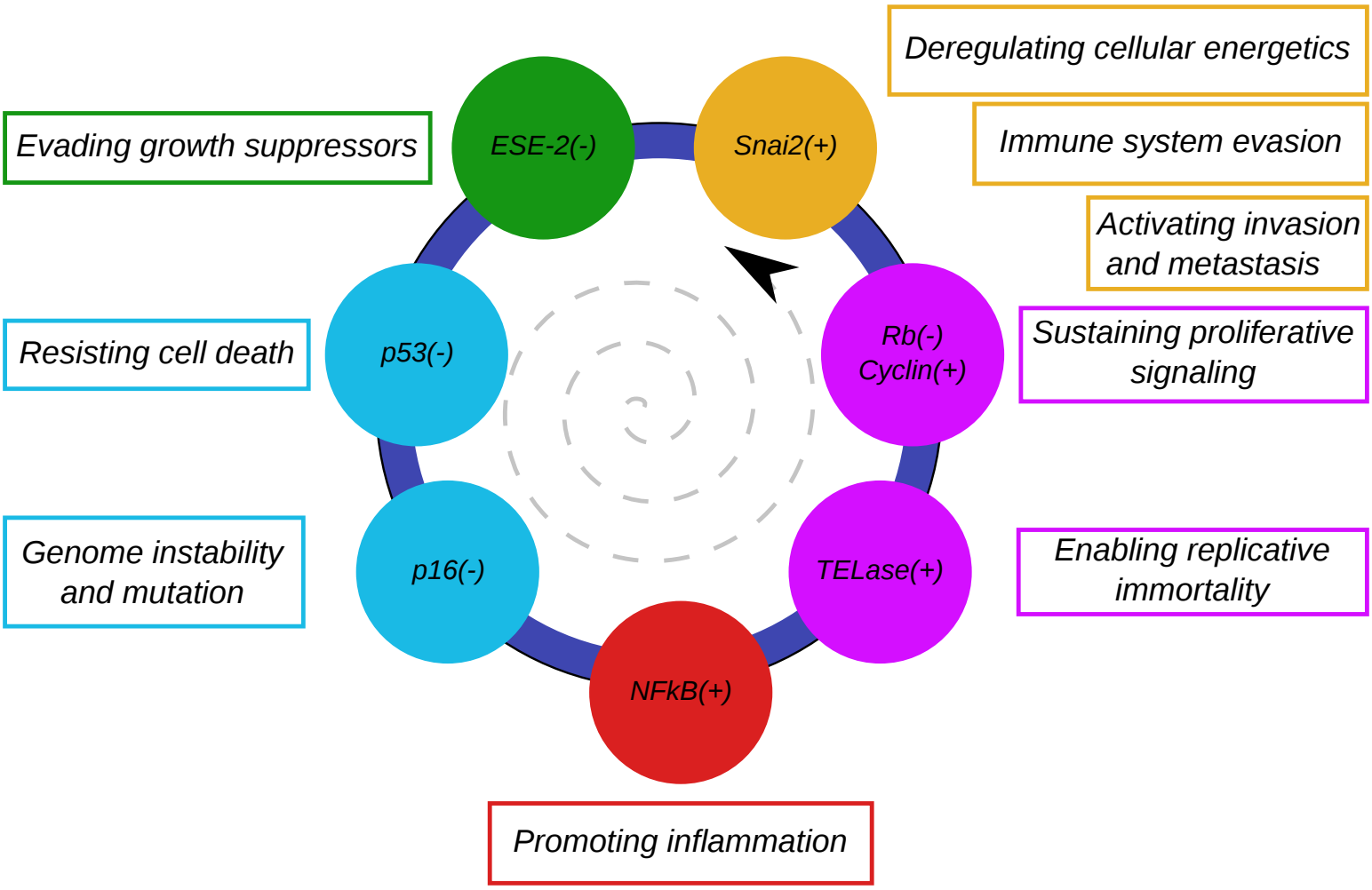
Table 1. Significant pathways and processes according to network–based gene set enrichment analysis

Cellular Phenotype	Recovered Attractor (Active)	*"Expected Attractors"*	References
Epithelial	Ese–2, NF-κB, E2F, Cyclin	Ese–2, NF–κB, Cell Cycle(+)	[29]
Senescent	p16, p53, Ese–2, NF–κB, Rb	p16, p53, NF–κB, Cell Cycle(-)	[42, 82, 83]
Mesenchymal stem-like	Snai2, Telomerase, NF–κB, Cyclin	Snai2, Telomerase, NF–κB, Cell Cycle(+)	[29, 42]

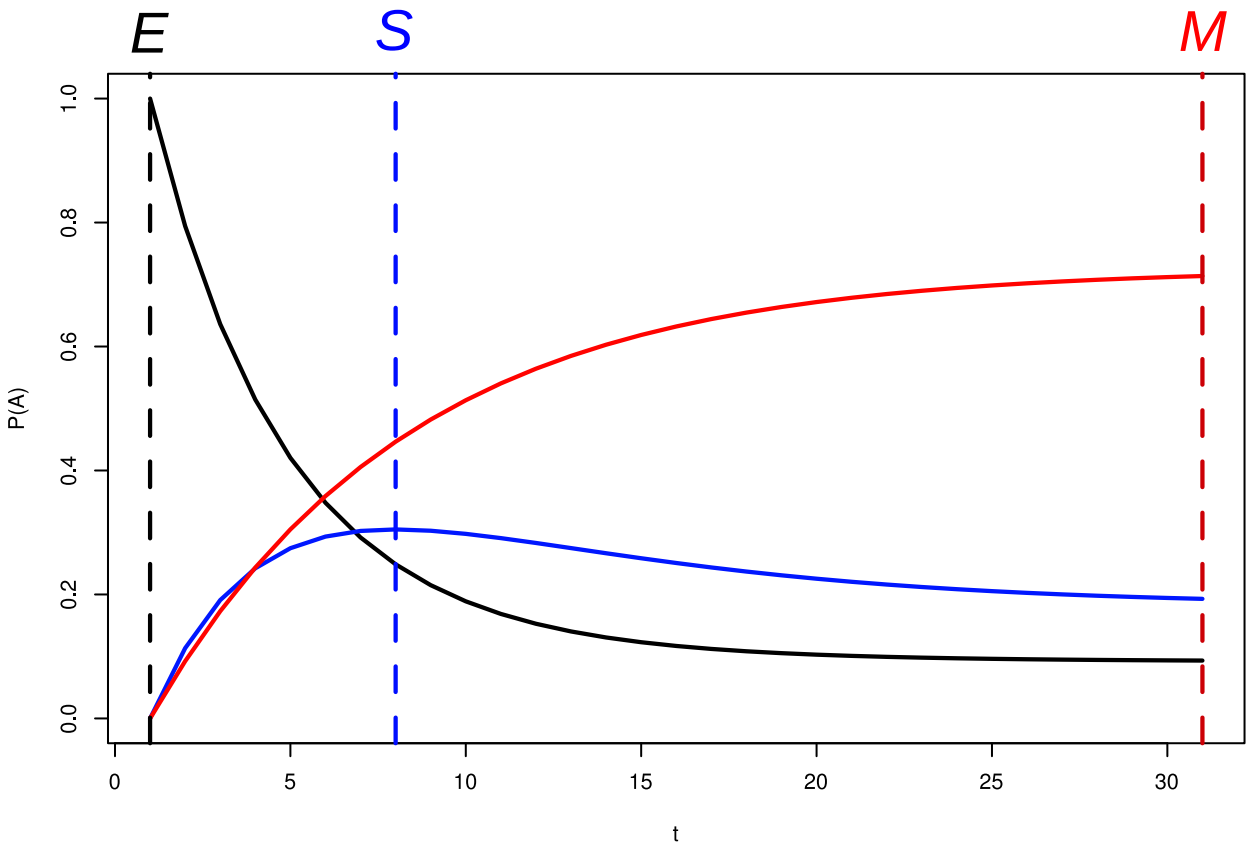
Table 2. Predicted and Observed Attractors


a)
TELun
TELasa
ATM
ATR
MDM2
Chk2
Chk1
p14
p21
p53
PML
Zeb
FOXC2
NF-kB
IkB
TNF-a
IL-6
TAK1
TGF-b
RKIP
EZH2
CDK6
CDK4
CDK2
CBX7
Snai2
Snail
Rb
Twist
ESE-2
ESE-3
ESE-1
HER2
Cyclin
SUZ12
p16
EED
BMI1
C-MYC
E2F
EGF
b)
TELun
TELasa
ATM
ATR
MDM2
Chk2
Chk1
p14
p21
p53
PML
Zeb
FOXC2
NF-kB
IkB
TNF-a
IL-6
TAK1
TGF-b
RKIP
EZH2
CDK6
CDK4
CDK2
CBX7
Snai2
Snail
Rb
Twist
ESE-2
ESE-3
ESE-1
HER2
Cyclin
SUZ12
p16
EED
BMI1
C-MYC
E2F
EGF

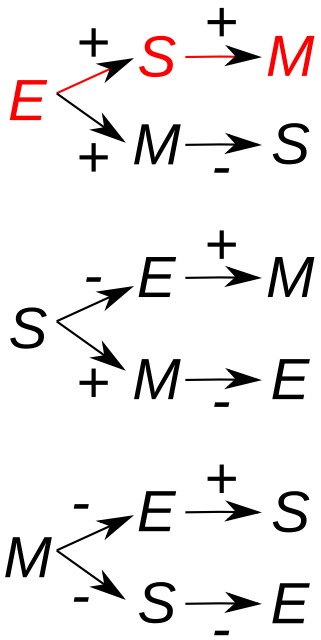

TELasa
p53
NF-kB
ESE-2
Snai2
Cyclin
p16
Rb
E2F

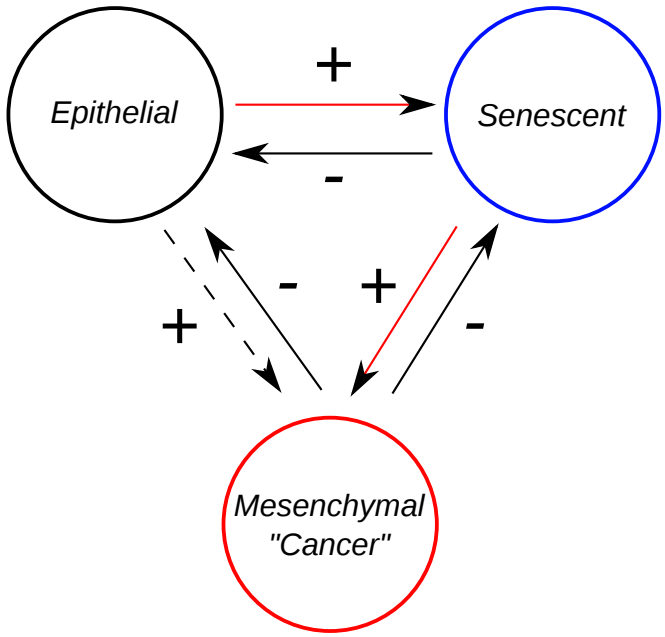

Deregulating cellular energetics
Immune system evasion
Activating invasion and metastasis
Sustaining proliferative signaling
Enabling replicative immortality
Promoting inflammation
Genome instability and mutation
Resisting cell death
Evading growth suppressors
ESE-2(-)
Snai2(+)
Rb(-)
Cyclin(+)
TELase(+)
NFkB(+)
p16(-)
p53(-)

a)


E
S
M
P(A)
0.0
0.2
0.4
0.6
0.8
1.0
0
5
10
15
20
25
30
t

b)


E
S
M
+
+
+
-
S
E
M
-
+
+
-
M
E
S
-
+
-
-


Epithelial
Senescent
Mesenchymal
"Cancer"
+
-
-
+
+
-

frontiers

Methods for Characterizing the Epigenetic Attractors Landscape Associated with Boolean Gene Regulatory Networks

Davila-Velderrain J [1,2,*], **Juarez-Ramiro L** [3], **Martinez-Garcia JC** [3], **and Alvarez-Buylla ER** [1,2,*]

[1] *Instituto de Ecología, Universidad Nacional Autónoma de México, Cd. Universitaria, México, D.F. 04510, México*
[2] *Centro de Ciencias de la Complejidad (C3), Universidad Nacional Autónoma de México, Cd. Universitaria, México, D.F. 04510, México*
[3] *Departamento de Control Automático, Instituto Politécnico Nacional, A. P. 14-740, 07300 México, DF, México*

Correspondence*:
Elena Alvarez-Buylla
Instituto de Ecología, Universidad Nacional Autónoma de México, Cd. Universitaria, México, D.F. 04510, México, eabuylla@gmail.com

Jose Davila-Velderrain
Instituto de Ecología, Universidad Nacional Autónoma de México, Cd. Universitaria, México, D.F. 04510, México, jdjosedavila@gmail.com

ABSTRACT

Gene regulatory network (GRN) modeling is a well established theoretical framework for the study of cell-fate specification during developmental processes. Recently, dynamical models of GRNs have been taken as a basis for formalizing the metaphorical model of Waddingtons epigenetic landscape, providing a natural extension for the general protocol of GRN modeling. In this contribution we present in a coherent framework a novel implementation of two previously proposed general frameworks for modeling the *Epigenetic Attractors Landscape* associated with boolean GRNs: the *inter-attractor* and *inter-state* transition approaches. We implement novel algorithms for estimating inter-attractor transition probabilities without necessarily depending on intensive single-event simulations. We analyze the performance and sensibility to parameter choices of the algorithms for estimating inter-attractor transition probabilities using three real GRN models. Additionally, we present for the first time, a side-by-side analysis of the two frameworks and show how the methods complement each other using a real case study: a cellular-level GRN model for epithelial carcinogenesis. We expect the toolkit and comparative analyzes put forward here to be a valuable additional resource for the systems biology community interested in modeling cellular differentiation and reprogramming both in normal and pathological developmental processes.

Keywords: Gene regulatory network, Epigenetic landscape, system dynamics, stochastic model, attractors, cell-fate decision, development

1 INTRODUCTION

The postulation of experimentally grounded gene regulatory network (GRN) dynamical models, their qualitative analysis and dynamical characterization in terms of control parameters, and the validation of GRN predictions against experimental observations has become a well established framework in systems biology – see, for example: **Mendoza and Alvarez-Buylla** (1998); **Espinosa-Soto et al.** (2004); **Huang et al.** (2007); **Davila-Velderrain et al.** (2015a). There are multiple tools available for the straightforward implementation and analysis of dynamical models of GRNs **Azpeitia et al.** (2014). These models are well-suited for the study of cell-fate specification during developmental processes. More recently, dynamical models of GRNs have been taken as a basis for formalizing a century-old developmental metaphor: Waddington's epigenetic landscape **Waddington** (1957); **Alvarez-Buylla et al.** (2008); **Huang** (2012); **Villarreal et al.** (2012); **Davila-Velderrain et al.** (2015c). The present authors recently introduced the term *Epigenetic Attractors Landscape (EAL)* in order to distinguish this modern view of the EL from its metaphorical counterpart (see **Davila-Velderrain et al.** (2015b)). Accordingly, here we will refer as EAL to a group of dynamical models grounded in dynamical systems theory and which operationally define an underlying EL associated with GRN dynamics. In this contribution we focus on the EAL associated with boolean GRNs.

Despite growing interest in modeling the EAL, as evidenced by recent model proposals in the study of stem cell differentiation **Li and Wang** (2013) and reprogramming **Wang et al.** (2014b), as well as the study of carcinogenesis **Wang et al.** (2014a); **Zhu et al.** (2015) and cancer therapeutics **Choi et al.** (2012); **Wang** (2013); unlike the case of GRNs, there are no available tools for the straightforward implementation of EALs. Furthermore, different EAL models have not been compared directly through side-by-side analysis of the same biological system. This has arguably precluded the wide-spread applicability of EALs.

One of the first methodological frameworks proposed to explore the EAL associated with a Boolean GRN was presented by Alvarez-Buylla and collaborators **Alvarez-Buylla et al.** (2008). Briefly, in its original form this framework rests on three steps: (1) introducing stochasticity into the boolean dynamics by means of the so-called stochasticity in nodes model (SIN), (2) estimating an *inter-attractor* transition probability matrix by simulation, and (3) analyzing the temporal evolution of the probability distribution over attractor states (see methods). For the purpose of this contribution, we refer to such framework as the *inter-attractor* transition approach (IAT). Recently, a related framework was presented by Zhou and collaborators **Zhou et al.** (2014a). The main differences between this and the former method are: the latter (1) precludes simulation by introducing stochasticity directly into a deterministic transition matrix, and (2) it is based on the estimation of a *inter-state* transition probability matrix. We refer to this latter framework as the *inter-state* transition approach (IST). Additionally, Zhou and collaborator introduced the idea of a global ordering of attractors in the EAL defined by analyzing the relative stability of attractor states **Zhou et al.** (2014b).

In this contribution we present in a coherent framework a novel implementation of the two methodologies, as well as associated analysis tools such as the global ordering of the attractors based on relative stabilities, the computation of a quasi-potential landscape based on an stationary probability distribution, and additional tools for downstream analyzes and plotting. We use the popular R statistical programming environment (www.R-project.org). For the first framework (IAT), we implement novel algorithms for estimating *inter-attractor* transition probabilities without necessarily depending on intensive single-event simulations. For both frameworks (IAT and IST) we exploit the vector-based programming capability of the R language. We analyze the performance and sensibility to parameter choices of the algorithms for estimating *inter-attractor* transition probabilities using three GRN models: the Arabidopsis (1) root stem cell niche and (2) early flower development GRNs; and a cellular-level GRN model for epithelial carcinogenesis. Additionally, for the latter model we present for the first time, a side-by-side analysis of the two frameworks and show how the methods complement each other. Importantly, we show that the attractor time-ordered transitions obtained by directly estimating an inter-attractor transition matrix are consistent with the global ordering of the attractors obtained by means of an

their corresponding relative stabilities. All the necessary codes for applying the methods showed herein are made publicly available; we expect this toolkit to be a valuable additional resource for the systems biology community.

2 RESULTS

2.1 CHARACTERIZING THE EPIGENETIC ATTRACTORS LANDSCAPE

In this work we organize previously existing, yet dispersed, mathematical analyzes into a coherent framework for the characterization of EAL associated with Boolean GRNs. Figure 1 schematically represents a general work flow for such characterization. The work flow is supposed to be applicable to an already available and validated experimentally grounded Boolean GRN model (see **Azpeitia et al.** (2014)). The first necessary step (Fig. 1a) consist on characterizing the state-space associated with the GRN in terms of the attained attractors and their basins, a standard practice in the dynamical analysis of Boolean GRNs (see methods). The second main step consists on estimating either a inter-attractor or inter-state transition probability matrix (or both) (Fig. 1b). The former is the main mathematical structure for the IAT aproach, and the latter for the IST approach (see methods). Downstream analyzes of the underlying EAL such as the temporal-order of attractor attainment, the attractor relative stability and global ordering, and the construction of a probabilistic landscape are based on the transition matrices and can be applied afterwards (Fig. 1c).

2.2 INTER-ATTRACTOR TRANSITIONS

A first necessary step in order to explore the EAL associated with a Boolean GRN using the IAT approach is to calculate the probabilities of transition from one attractor to another. In this contribution we present two algorithms for such task (see methods). Algorithm 1 implements what we will refer to as an intuitive mapping-guided random walk in state space. The reasoning is as follows. An initial state is taken at random, which is then mapped to a next state using the stochastic mapping in Equation (3). The basins corresponding to the two states are recorded in order. Subsequently, another state is picked at random from the latter basin and the mapping procedure is repeated. The procedure is repeated $Nsteps$ number of times, each time taking at random a state from the present basin, and the goal is to record a stochastic realization of the transitions from one basin to another. Algorithm 2, on the other hand, considers all the possible states, repeats them $Nreps$ number of times in a single data structure, and maps them using Equation (3) as well (for details, see methods). An important technical issue is then how to select the parameters $Nsteps$ and $Nreps$, respectively. Specially, because this type of simulation approaches have been qualified as requiring large number of time-consuming sampling **Zhou et al.** (2014a).

For each algorithm we tested how the estimate of the inter-attractor transition matrix changes as the parameter value increases. We used three real GRN models for testing: *Arabidopsis* single-cell root stem cell niche GRN (root-GRN) **Azpeitia et al.** (2010), *Arabidopsis* floral organ determination GRN (flower-GRN) **Azpeitia et al.** (2014), and a cellular-level GRN model for epithelial carcinogenesis (cancer-GRN). We found that for models of size common to GRN developmental modules (i.e., $8 - 15$ genes) the estimation obtained with small values of the parameter rapidly converges to that obtained by using large values (e.g., $\approx 10^6$). Figure 2 shows how the distance between the estimate obtained using a value $Nsteps(Nreps) = i$ and that obtained using $Nsteps = 10^6$ and $Nreps = 10^3$ for Algorithms 1 and 2, respectively. These results correspond to the three GRN models: root (Fig. 2a-b), cancer (Fig. 2c-d), and flower (Fig. 2e-f). Additionally, we show that the estimate obtained with one of the algorithms also rapidly converges to that obtained with the other algorithm. Figure 3 shows how the distance between the estimate obtained using one algorithm with a parameter value i and that obtained using the other algorithm with a large parameter value decreases as i increases. Based on this latter analysis we conclude that, for GRNs of sizes $8-15$ genes, using a value of the order of $Nsteps = 10^4$ for algorithm 1 and $Nreps = 10^2$

would be sufficient to achieve an accuracy similar to that achieved using large values (i.e, 10^6 and 10^3, respectively).

2.3 CHARACTERIZING THE EAL

In this section we provide as an example the analysis of the EAL underlying a cellular-level GRN model for epithelial carcinogenesis. The details of the construction and validation of such network model are being published by the authors elsewhere. The GRN comprises 9 main regulators of epithelial carcinogenesis (Fig. 4), and its dynamical characterization uncovers 3 fixed-point attractor corresponding to the epithelial, senescent, and messenchymal stem-like cellular phenotypes. We applied the two approaches (IAT and IST) to the cancer-GRN, and for the IAT approach we applied the two algorithms proposed herein. Accordingly, we estimated two inter-attractor transition matrices and one inter-state transition matrix. For simplicity in all cases we kept fixed a single value for the error parameter $\xi = 0.05$. Using the estimated matrices, we applied the downstream analyzes depicted in Figure 1c. Figure 5 shows two graphs plotting the temporal evolution of the occupation probability distribution over attractor states epithelial (black), senescent (red) and messenchymal (green) – conditioned on an initial distribution where all the cellular population is in the epithelial attractor state. The uncovered attractor time-order is indicated by sequential vertical lines: the order is epithelial $\rightarrow$ senescent $\rightarrow$ messenchymal. Importantly, the two algorithms give the same qualitative result.

Subsequently, we uncovered the global ordering of attractors by calculating the relative stabilities and net transition rates between pairs of attractors using the two inter-attractor transitions estimated with the two algorithms (for details, see methods). Figure 6 shows the plot of two graphs where an arrow appears in color red if the calculated transition rate between the attractor is positive in the indicated direction. The global ordering corresponds to the path comprised by directed arrows passing by the three attractors, here: epithelial $\rightarrow$ senescent $\rightarrow$ messenchymal. Thus, the global ordering is consistent with the attractor time-order, as long as the latter is conditioned on having the total probability mass in the epithelial attractor as initial state. Again, the two algorithms produce the same qualitative result.

Finally, we used the estimated inter-state transition matrix obtained with the IST approach to derive a graphical probabilistic landscape (see methods). The landscape is based on the stationary probability distribution $\mathbf{u}_{ss}$ obtained by numerical simulation (see methods). Figure 7 and 8 show a 3D-surface and a contour plot respectively. The graphical landscape was derived by first mapping all the state vectors in the sate-space into a low dimensional space by the dimensionality reduction technique principal component analysis. The first two component are taken as the coordinates in the 3D plot, where the z-coordinate corresponds to the values $-log(\mathbf{u}_{ss})$. The surface is inferred by interpolating the spaced data points using the technique of thin plate spline regression **Furrer et al.** (2009). The 3D-surface plot nicely shows the relative stability of the states by means of their probability, the lower states begin more stable. The route from the attractors of less stability to that with the highest consists with the global ordering uncovered above. However, in the case of the IST transition and the probabilistic landscape we have additional information concerning the relative stability of all the transitory states in state space.

3 DISCUSSION

Boolean GRN models are well-established tools for the mechanistic study of the establishment of cellular phenotypes during developmental dynamics. Their simplicity and deterministic nature are well-suited for answering questions regarding the sufficiency of molecular players and interactions necessary to explain observed cellular phenotypes. In the present contribution we present methods to study an extended Boolean GRN model which take stochasticity into consideration, necessary for studying cell-state transition events.

In the case of the stochastic Boolean GRNs, the model of interest involves random samples with a non-trivial dependence structure. In such cases, efficient simulation algorithms are needed in order to explore

and characterize the underlying structure and to understand the behavioral (dynamical) consequences of the constrains imposed by such structure. Accordingly, we propose two algorithms of general applicability, and show how these can be used to estimate transitions probabilities in an efficient way from moderate size GRNs similar to those proposed as developmental modules driving developmental processes. Although we show that the two algorithms generate consistent estimates, one or the order may be prefer depending on the GRN in question and the computational resources at hand. Algorithm 1 is likely to be preferred in the case of larger GRNs, as it is not constrained by the size of the GRN per se, but the number of steps chosen in the simulation. On the other hand, given the declarative representation used in algorithm 2, its performance is constrained by the available of memory. Algorithm 2, however, may be preferred for fast estimates in small to moderate size GRNs (¡15 genes). Importantly, although we tested the performance of the algorithms in terms of the number of steps chosen for the simulations, the results should not be generalized without caution given that we only used three real GRNs, and the results may vary either for larger GRNs or sate spaces with more complex structures.

For illustrative purposes we applied all the methods and downstream analyzes presented herein to a specific GRN: a cellular-level GRN model for epithelial carcinogenesis. We show that for this case, the uncovered temporal-order of attractor attainment is consistent with the global ordering based on relative stability, both calculated from a inter-attractor transition probability matrix. The result of the former is conditioned on the initial occupation probability taken. An interesting open problem would be to generalize this relationship using GRNs with divers structures, for example to ask if the global ordering of attractors is robust enough as to drive most initial distributions into a consistent temporal ordering. An additional interesting questions would be, what does this relationship tells us about the structural constraints imposed by the GRN. The tools and implementation presented here may prove useful for such theoretical studies.

Finally, we present tools for deriving a probabilistic landscape from an estimated inter-state transition matrix in terms of the stationary probability distribution over state space. This latter analysis and the associated graphical tools can be applied to systematically study how the system responds to perturbations resulting in a reshaped EAL. Structural alterations of the EAL may predict the induction of preferential cell-state transitions such as the case of reprogramming strategies **Zhou and Huang** (2011) or therapeutic interventions against the stabilization of a cancer attractor **Huang and Kauffman** (2013); **Wang** (2013).

Overall, in this contribution we present in a coherent framework a novel implementation of general frameworks for modeling the *Epigenetic Attractors Landscape* associated with boolean GRNs. We provide analysis of the method performance and show how they can be applied to real case GRNs. we expect the toolkit and comparative analyzes put forward here to be a valuable additional resource for the systems biology community interested in modeling cellular differentiation and reprogramming both in normal and pathological developmental processes.

4 MATERIAL & METHODS

BOOLEAN GENE REGULATORY NETWORKS

A Boolean network models a dynamical system assuming both discrete time and discrete state variables. This is expressed formally with the mapping:

$$x_i(t+1) = F_i(x_1(t), x_2(t), ..., x_k(t)), \tag{1}$$

where the set of functions F_i are logical prepositions (or truth tables) expressing the relationship between the genes that share regulatory interactions with the gene i, and where the state variables $x_i(t)$ can take the discrete values 1 or 0 indicating whether the gene i is expressed or not at a certain time t, respectively.

A completely specified Boolean GRN model is analyzed by either of two methods: (1) by exhaustive computational characterization of the state space in terms of attained attractors and their basins of attractions (used in IAT), or (2) by defining a matrix explicitly encoding the mapping in Equation (1) (used in IST). Specifically, for the latter method, following **Zhou et al.** (2014b) the mapping in Equation (1) is used to define a single-step 2^n x 2^n transition matrix $\mathbf{T}$ with elements $t_{i,j}$, where:

$$t_{i,j} = \begin{cases} 1, & \mathbf{x}_j = \mathbf{F}(\mathbf{x}_i) \\ 0, & Otherwise. \end{cases} \tag{2}$$

Here $\mathbf{x}_i$ is the network state i from the state-space of size 2^n corresponding to a network of n genes, and $\mathbf{F}$ represents the vector of n functions represented element-wise in Equation (1). Given the deterministic character of the mapping in Equation (1), the matrix $\mathbf{T}$ is sparse, each row i having only one element where $t_{i,j} = 1$. The matrix $\mathbf{T}$ constitutes a declarative representation which includes the complete information of the mapping in Equation (1): the matrix $\mathbf{T}$ assign to each of the states $\mathbf{x}_k$, where $k \in \{1, ..., 2^n\}$, its corresponding state in time $t + 1$.

INTER-ATTRACTOR TRANSITION APPROACH

Including Stochasticity

Following **Alvarez-Buylla et al.** (2008); **Azpeitia et al.** (2014); **Davila-Velderrain et al.** (2015b), a Boolean GRN is extended into a discrete stochastic model by means of the so–called stochasticity in nodes (SIN) model. In this model, a constant probability of error ξ is introduced for the deterministic Boolean functions as follows:

$$\begin{aligned} P_{x_i(t+1)}[F_i(\mathbf{x}_{reg_i}(t))] &= 1 - \xi, \\ P_{x_i(t+1)}[1 - F_i(\mathbf{x}_{reg_i}(t))] &= \xi. \end{aligned} \tag{3}$$

It is assumed that the probability that the value of the random variable $x_i(t + 1)$ (a gene) is determined or not by its associated logical function $F_i(\mathbf{x}_{reg_i}(t))$ is $1-\xi$ or ξ, respectively. The probability ξ is a scalar constant parameter acting independently per gene. The vector $\mathbf{x}_{reg_i}$ represents the regulators of gene i.

Inter-Attractor Transition Probability Estimation

An attractor transition probability matrix Π with components:

$$\pi_{ij} = P(A_{t+1} = j | A_t = i), \tag{4}$$

representing the probability that an attractor j is reached from an attractor i is estimated by either of two simulation-based algorithms proposed herein (see results).

In Algorithm 2, $Bin(n = 1, \xi)$ refers to a binomial distribution given by $Bin(k|n, \xi) = \binom{n}{k} \xi^k (1 - \xi)^{n-k}$. In the special case used here (with $n = 1$) the distribution corresponds to a Bernoulli distribution. Thus, what we call *perturbation indicator vector* effectively simulates tossing a biased coin $Nsteps \,\mathrm{x}\, n \,\mathrm{x}\, 2^n$ times. Each outcome $x = 1$ indicates the position where an error in the mapping has occurred, according to Equation (3).

The elements π_{ij} of the matrix Π are obtained as maximum likelihood estimates based on the empirical transition probability resulting from the simulations from either algorithm 1 or 2.

Algorithm 1 Simulate *inter-attractor* stochastic realization

Initiate storage[*Nsteps*]
from $state\ space = \{1, ..., 2^n\}$ pick randomly initial state $\mathbf{x}_i$
storage[1] $\leftarrow$ basin $k \leftarrow$ map $\leftarrow \mathbf{x}_i$
for (*stepN* in 2 to *Nsteps*) **do**
  state $\mathbf{x}_j \leftarrow$ stochastic mapping Eq(2) $\leftarrow$ state $\mathbf{x}_i$
  storage[stepN] $\leftarrow$ basin $k \leftarrow$ map $\leftarrow \mathbf{x}_j$
  from $sub\ space = \{basin\ k\}$ pick randomly state $\mathbf{x}_i$
end for
return storage

Algorithm 2 Implicit bit-flip simulation

Initiate storage $j \times j$ matrix Π, $j \in \{1, ..., n_{attractors}\}$
Generate $state\ space = \{\mathbf{x}_1, ..., \mathbf{x}_{2^n}\}$
Generate set $\mathbf{X_{t+1}} = \mathbf{F}(state\ space)$
$\mathbf{X^{pert}_{t+1}} \leftarrow$ repeat $\mathbf{X_{t+1}}$ element-wise $Nsteps$ times
Generate perturbation indicator vector $\mathbf{piv}$:
  $\mathbf{piv} \leftarrow$ simulate $Nsteps \times n \times 2^n$ observations from $Bin(n = 1, \xi)$
for $\mathbf{piv}[i] = 1$ **do**
  Apply error in $\mathbf{X^{pert}_{t+1}}[i]$, $i \in \{1, ..., Nsteps \times n \times 2^n\}$
end for
$\mathbf{X^{pert}} \leftarrow$ split $\mathbf{X^{pert}_{t+1}}$ in n-size state vectors $\mathbf{x}_k, k \in \{1, ..., Nsteps \times 2^n\}$
for each $\mathbf{x}_i$ in $state\ space$ **do**
  basin $j \leftarrow$ map $\mathbf{x}_i$
end for
for each $\mathbf{x}_k$ in $\mathbf{X^{pert}}$ **do**
  basin $j \leftarrow$ map $\mathbf{x}_k$
end for
update $\pi_{j,j}$
return matrix

INTER-STATE TRANSITION PROBABILITY APPROACH

Including Stochasticity

For the IST approach, stochasticity is introduced in a declaractive manner (i.e., by means of a single structure representation) using a binomial distribution **Zhou et al.** (2014a,b). Specifically, the effect of noise on each possible single-state transition is represented by introducing a noise matrix $\mathbf{N}$ with elements

$$N_{i,j} = \begin{cases} \binom{n}{d_{ij}} \xi^{d_{ij}} (1-\xi)^{n-d_{ij}}, & i \neq j \\ 0, & i = j \end{cases} \tag{5}$$

where d_{ij} is the Hamming distance between the states i and j (i.e., $d_{ij} = \|\mathbf{x}_i - \mathbf{x}_j\|_H$). This representation formalizes an intuitive notion: the effect of noise on the system is more (less) likely to produce a state less (more) similar to the initial state.

Inter-State Transition Probability Estimation

A single object including both stochastic perturbations and deterministic mapping is obtained by adding the noise matrix $\mathbf{N}$ and the deterministic single-step transition matrix $\mathbf{T}$ (see Equation 2) as follows

$$\mathbf{\Pi} = (1-\xi)^n \mathbf{T} + \mathbf{N} \tag{6}$$

After normalizing a transition probability matrix Π is obtained with components

$$\pi_{ij} = P(\mathbf{x}_{t+1} = j | \mathbf{x}_t = i). \tag{7}$$

The components π_{ij} represent the probability that a state j is reached from a state i, where $i, j \in \{1, ...2^n\}$.

TEMPORAL EVOLUTION OF STATES/ATTRACTORS PROBABILITY

In both approaches (IAT and IST) a sequence of random variables $\{C_t : t \in \mathbb{N}\}$ is considered a Markov chain (MC). In IAT (IST) C_T takes as values the different attractors (states), the elements $\pi_{i,j}$ representing inter-attactor(states) transition probabilities, and the matrix Π the (one-step) transition probability matrix. As the probabilities do not depend on time, the MC is homogeneous.

The occupation probability distribution $P(C_t = j)$ – i.e., the probability that the chain is in state (attractor or state) j at a given time t – is denoted by the row vector $\mathbf{u}(t)$. The probabilities temporally evolve according to the dynamic equation

$$\mathbf{u}(t+1) = \mathbf{u}(t)\mathbf{\Pi}. \tag{8}$$

Taking $\mathbf{u}(0)$ as the initial distribution of the MC, the equation reads $\mathbf{u}(1) = \mathbf{u}(0)\mathbf{\Pi}$. By linking the occupation probabilities iteratively we get $\mathbf{u}(t) = \mathbf{u}(0)\mathbf{\Pi}^t$: the occupation probability distribution at time t can be obtained directly by matrix exponentiation.

EAL ANALYZES

Temporal-order of Attractor Attainment

Having obtained the temporal evolution of the occupation probability distribution $\mathbf{u}(t)$ given an initial distribution $\mathbf{u}(0)$ by numerically solving Equation (8), following **Alvarez-Buylla et al.** (2008), it is assumed that the most likely time for an attractor to be reached is when the probability of reaching that particular attractor is maximal. Therefore, the temporal sequence in which attractors are attained is obtained by determining the sequence in which their maximum probabilities are reached using $\mathbf{u}(t)$.

Probabilistic Landscape

A stationary probability distribution of a MC is a distribution $\mathbf{u}_{ss}$ which satisfies the steady state equation $\mathbf{u}_{ss} = \mathbf{u}_{ss}\mathbf{\Pi}$. The stationary probability distribution, if exists, is calculated either by solving the equation $\mathbf{u}_{ss}(\mathbf{I} - \mathbf{\Pi}) = 0$, where $\mathbf{I}$ is the $n\,x\,n$ identity matrix **Wilkinson** (2011); or by numerically solving Equation (8), as $\mathbf{u}_{ss}$ corresponds to the *long-run distribution* of the MC: $\mathbf{u}_{ss} = \lim_{t\to\infty} \mathbf{u}(t)$ **Bolstad** (2011). A probabilistic landscape U – also called a quasi-potential – can be obtaining by mapping the distribution $\mathbf{u}_{ss}$ using $-ln(\mathbf{u}_{ss})$. Such landscape reflects the probability of states and it provides a global characterization and a stability measure of the GRN system **Wang** (2015).

Attractor Relative Stability and Global Ordering Analyses

A relative stability matrix M is calculated which reflects the transition barrier between any two states based on the mean first passage time (MFPT). The transition barrier in the EAL epitomizes the ease for transitioning from one attractor to another. The ease of transitions, in turn, offers a notion of relative stability. Zhou and collaborators recently proposed that a GRN has a consistent global ordering of all of the attractors which can be uncovered by considering their relative stabilities **Zhou et al.** (2014a,b). A net transition rate between attractor i and j is defined in terms of the MFPT as follows:

$$d_{i,j} = \frac{1}{MFPT_{i,j}} - \frac{1}{MFPT_{j,i}} \tag{9}$$

The consistent global ordering of the attractors is defined based on the formula proposed in **Zhou et al.** (2014b). Briefly, the consistent global ordering of the attractors is given by the attractor permutation in which al transitory net transition rates from an initial attractor to a final attractor are positive. The MFPTs are calculated either by implementing the matrix-based algorithm proposed in) or by means of numerical simulation.

IMPLEMENTATION

All the methods presented here were implemented using the *R* statistical programming environment (www.R-project.org). The code relies on the following packages: *BoolNet*, for the dynamical analysis oo Boolean neworks **Müssel et al.** (2010); *expm*, for matrix computations **Goulet et al.** (2013); *igraph*, for network analyses **Csardi and Nepusz** (2006); *markovchain* for MC analysis and inference; and *fields*, for surface plotting **Furrer et al.** (2009).

DISCLOSURE/CONFLICT-OF-INTEREST STATEMENT

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

REFERENCES

Alvarez-Buylla, E. R., Chaos, Á., Aldana, M., Benítez, M., Cortes-Poza, Y., Espinosa-Soto, C., et al. (2008), Floral morphogenesis: stochastic explorations of a gene network epigenetic landscape, *Plos one*, 3, 11, e3626

Azpeitia, E., Benítez, M., Vega, I., Villarreal, C., and Alvarez-Buylla, E. R. (2010), Single-cell and coupled grn models of cell patterning in the arabidopsis thaliana root stem cell niche, *BMC systems biology*, 4, 1, 134

Azpeitia, E., Davila-Velderrain, J., Villarreal, C., and Alvarez-Buylla, E. R. (2014), Gene regulatory network models for floral organ determination, in Flower Development: Methods and Protocols (Springer)

Bolstad, W. M. (2011), Understanding computational Bayesian statistics, volume 644 (John Wiley & Sons)

Choi, M., Shi, J., Jung, S. H., Chen, X., and Cho, K.-H. (2012), Attractor landscape analysis reveals feedback loops in the p53 network that control the cellular response to dna damage, *Science signaling*, 5, 251, ra83–ra83

Csardi, G. and Nepusz, T. (2006), The igraph software package for complex network research, *InterJournal, Complex Systems*, 1695, 5, 1–9

Davila-Velderrain, J., Martinez-Garcia, J., and Alvarez-Buylla, E. (2015a), Descriptive vs. mechanistic network models in plant development in the post-genomic era, *Plant Functional Genomics: Methods and Protocols*, 455–479

Davila-Velderrain, J., Martinez-Garcia, J. C., and Alvarez-Buylla, E. R. (2015b), Modeling the epigenetic attractors landscape: toward a post-genomic mechanistic understanding of development, *Frontiers in genetics*, 6

Davila-Velderrain, J., Villarreal, C., and Alvarez-Buylla, E. R. (2015c), Reshaping the epigenetic landscape during early flower development: induction of attractor transitions by relative differences in gene decay rates, *BMC systems biology*, 9, 1, 20

Espinosa-Soto, C., Padilla-Longoria, P., and Alvarez-Buylla, E. R. (2004), A gene regulatory network model for cell-fate determination during arabidopsis thaliana flower development that is robust and recovers experimental gene expression profiles, *The Plant Cell Online*, 16, 11, 2923–2939

Furrer, R., Nychka, D., and Sain, S. (2009), fields: Tools for spatial data, *R package version*, 6, 11

Goulet, V., Dutang, C., Maechler, M., Firth, D., Shapira, M., and Stadelmann, M. (2013), expm: Matrix exponential, *R package version 0.99-0*

Huang, S. (2012), The molecular and mathematical basis of waddington's epigenetic landscape: A framework for post-darwinian biology?, *Bioessays*, 34, 2, 149–157

Huang, S., Guo, Y.-P., May, G., and Enver, T. (2007), Bifurcation dynamics in lineage-commitment in bipotent progenitor cells, *Developmental biology*, 305, 2, 695–713

Huang, S. and Kauffman, S. (2013), How to escape the cancer attractor: rationale and limitations of multi-target drugs, in Seminars in cancer biology, volume 23 (Elsevier), volume 23, 270–278

Li, C. and Wang, J. (2013), Quantifying cell fate decisions for differentiation and reprogramming of a human stem cell network: landscape and biological paths, *PLoS computational biology*, 9, 8, e1003165

Mendoza, L. and Alvarez-Buylla, E. R. (1998), Dynamics of the genetic regulatory network for arabidopsis thaliana flower morphogenesis, *Journal of theoretical biology*, 193, 2, 307–319

Müssel, C., Hopfensitz, M., and Kestler, H. A. (2010), Boolnetan r package for generation, reconstruction and analysis of boolean networks, *Bioinformatics*, 26, 10, 1378–1380

Villarreal, C., Padilla-Longoria, P., and Alvarez-Buylla, E. R. (2012), General theory of genotype to phenotype mapping: derivation of epigenetic landscapes from N-node complex gene regulatory networks., *Physical review letters*, 109, 11, 118102

Waddington, C. H. (1957), The strategy of genes (London: George Allen & Unwin, Ltd.)

Wang, G., Zhu, X., Gu, J., and Ao, P. (2014a), Quantitative implementation of the endogenous molecular–cellular network hypothesis in hepatocellular carcinoma, *Interface focus*, 4, 3, 20130064

Wang, J. (2015), Landscape and flux theory of non-equilibrium dynamical systems with application to biology, *Advances in Physics*, 64, 1, 1–137

Wang, P., Song, C., Zhang, H., Wu, Z., Tian, X.-J., and Xing, J. (2014b), Epigenetic state network approach for describing cell phenotypic transitions, *Interface Focus*, 4, 3, 20130068

Wang, W. (2013), Therapeutic hints from analyzing the attractor landscape of the p53 regulatory circuit, *Science signaling*, 6, 261, pe5–pe5

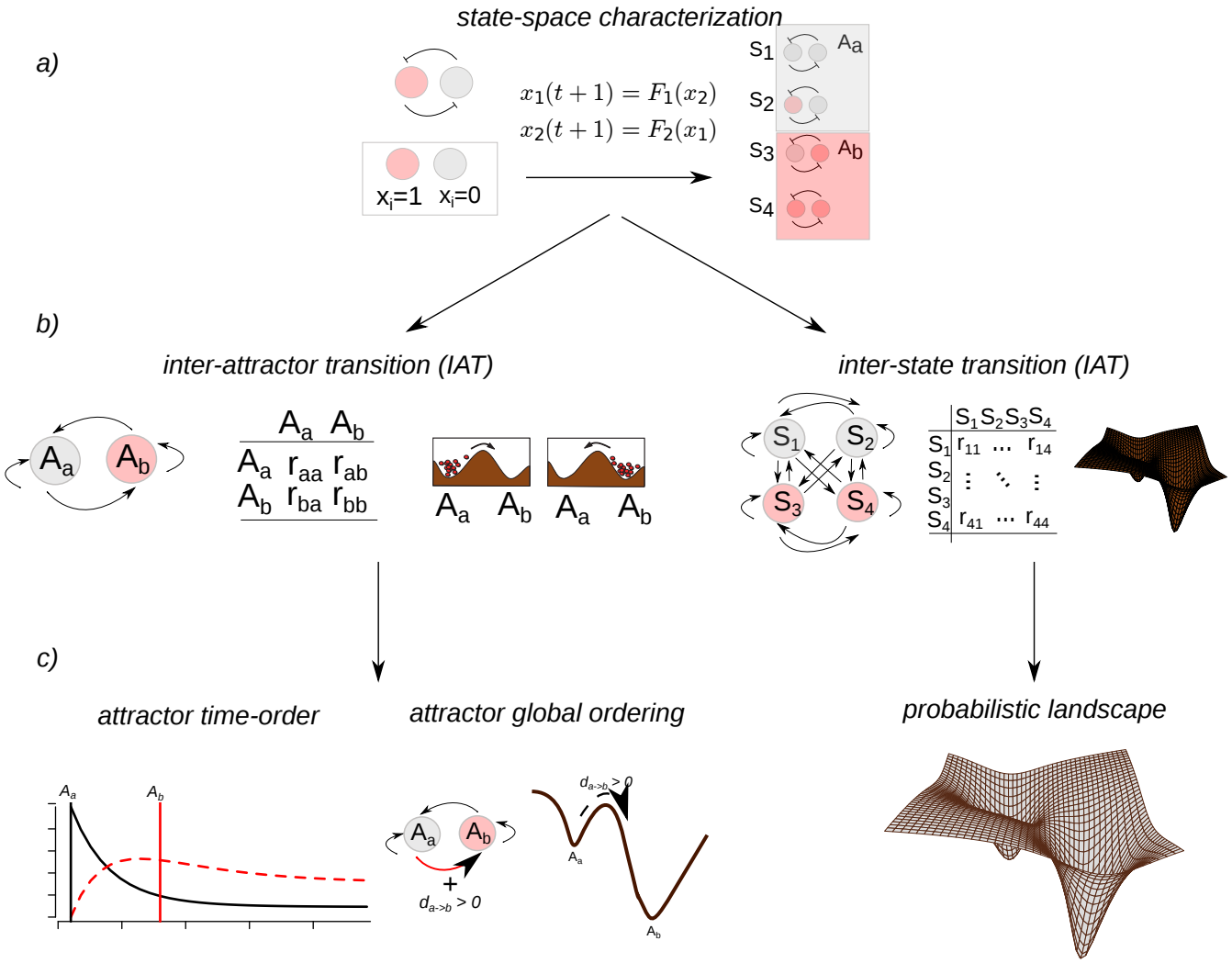
Wilkinson, D. J. (2011), Stochastic modelling for systems biology (CRC press)

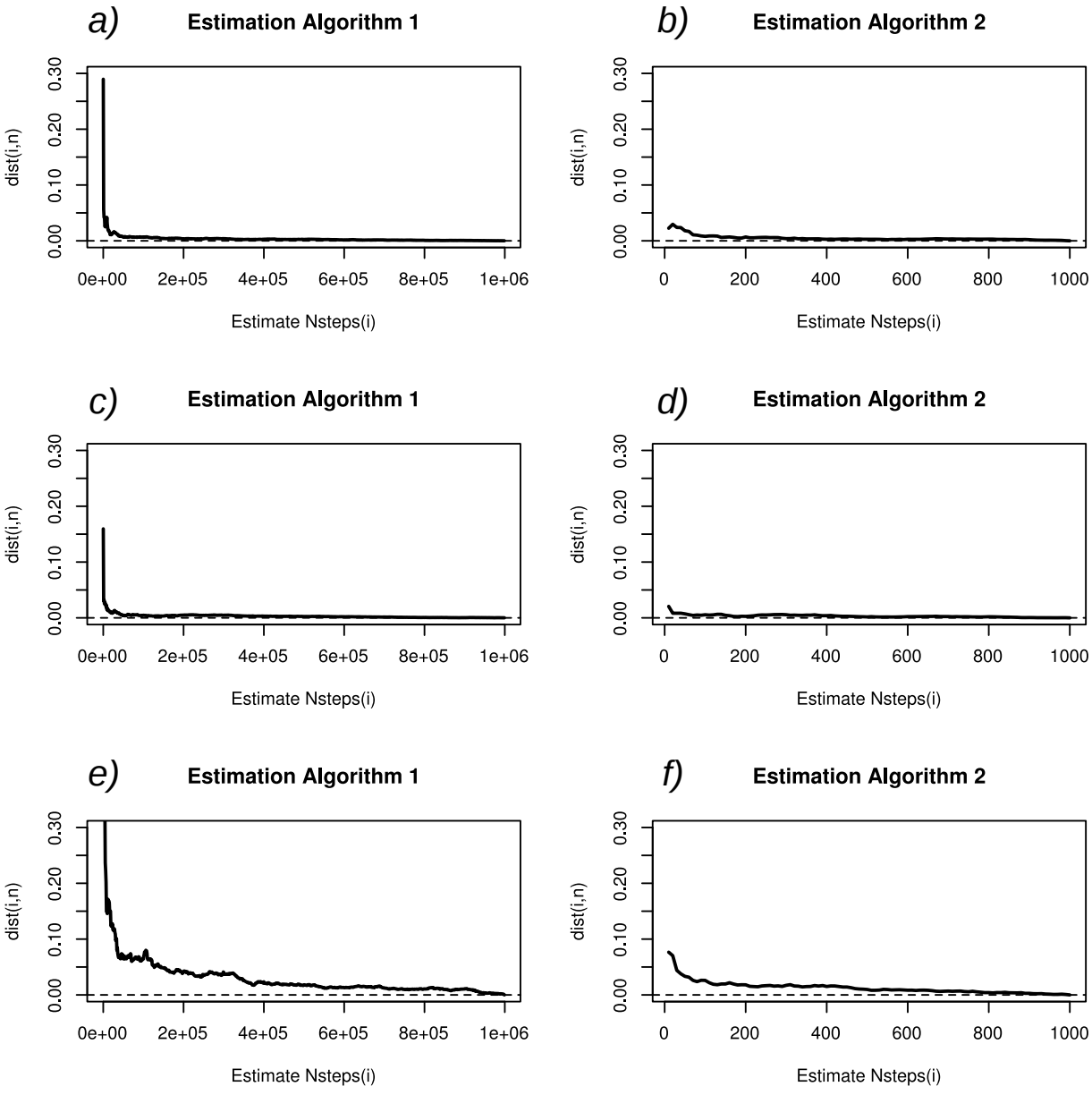
Zhou, J. X. and Huang, S. (2011), Understanding gene circuits at cell-fate branch points for rational cell reprogramming, *Trends in Genetics*, 27, 2, 55–62

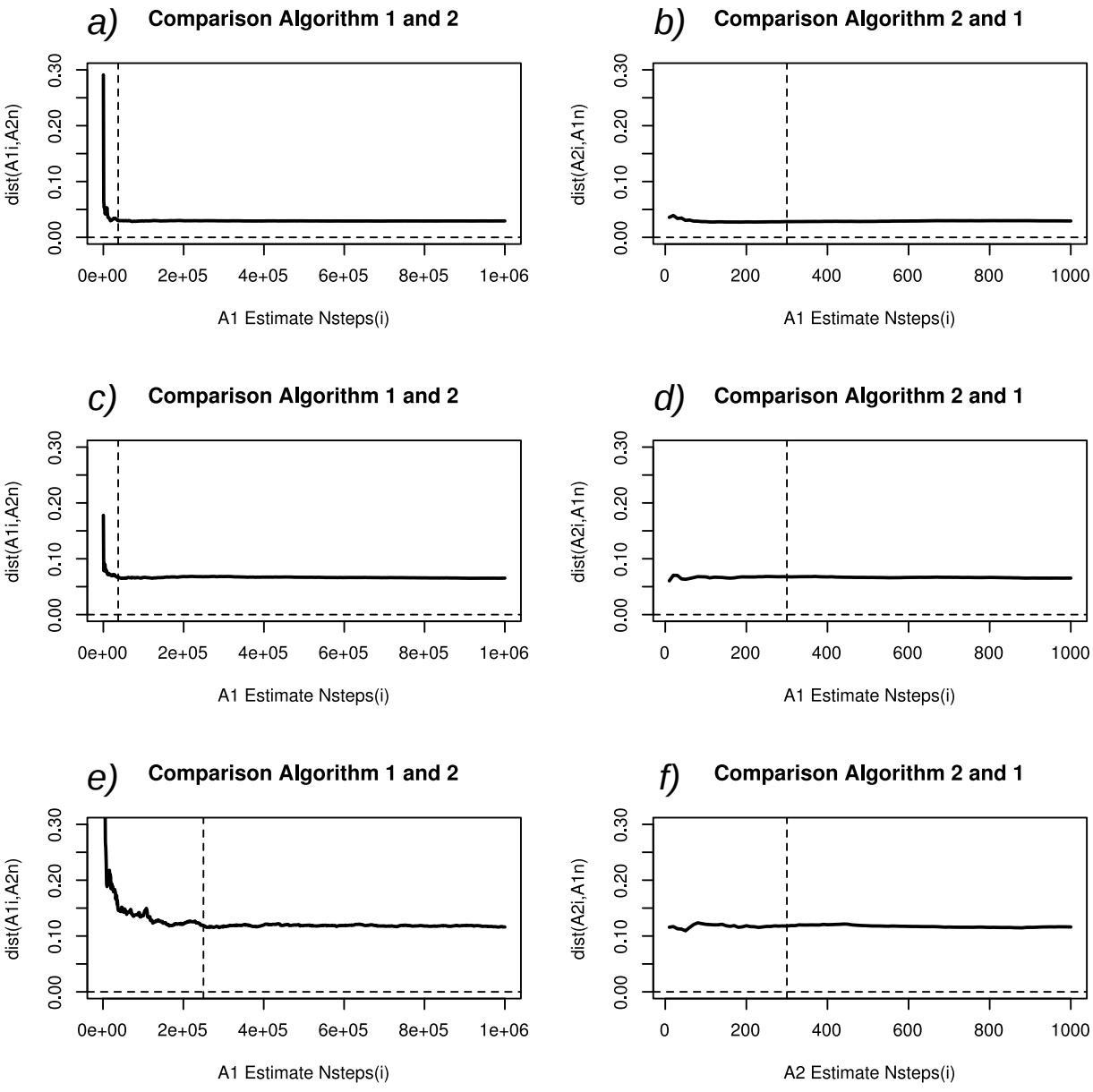
Zhou, J. X., Qiu, X., d'Herouel, A. F., and Huang, S. (2014a), Discrete gene network models for understanding multicellularity and cell reprogramming: From network structure to attractor landscapes landscape, *In: Computational Systems Biology. Second Edition. Elsevier*, 241–276

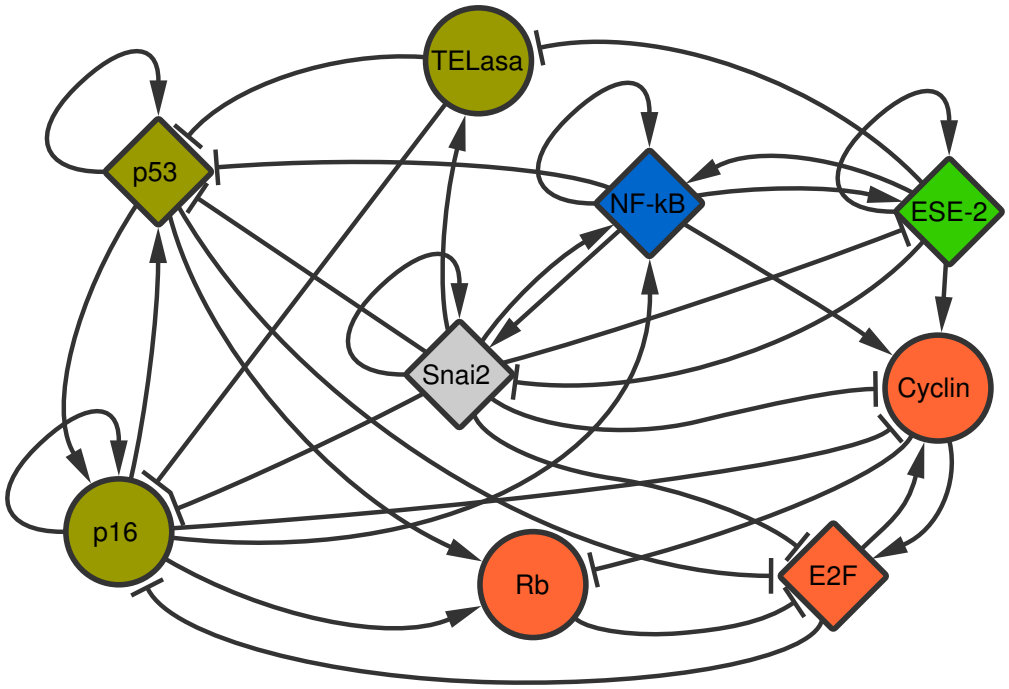
Zhou, J. X., Samal, A., d'Hèrouël, A. F., Price, N. D., and Huang, S. (2014b), Relative stability of network states in boolean network models of gene regulation in development, *arXiv preprint arXiv:1407.6117*

Zhu, X., Yuan, R., Hood, L., and Ao, P. (2015), Endogenous molecular-cellular hierarchical modeling of prostate carcinogenesis uncovers robust structure, *Progress in biophysics and molecular biology*

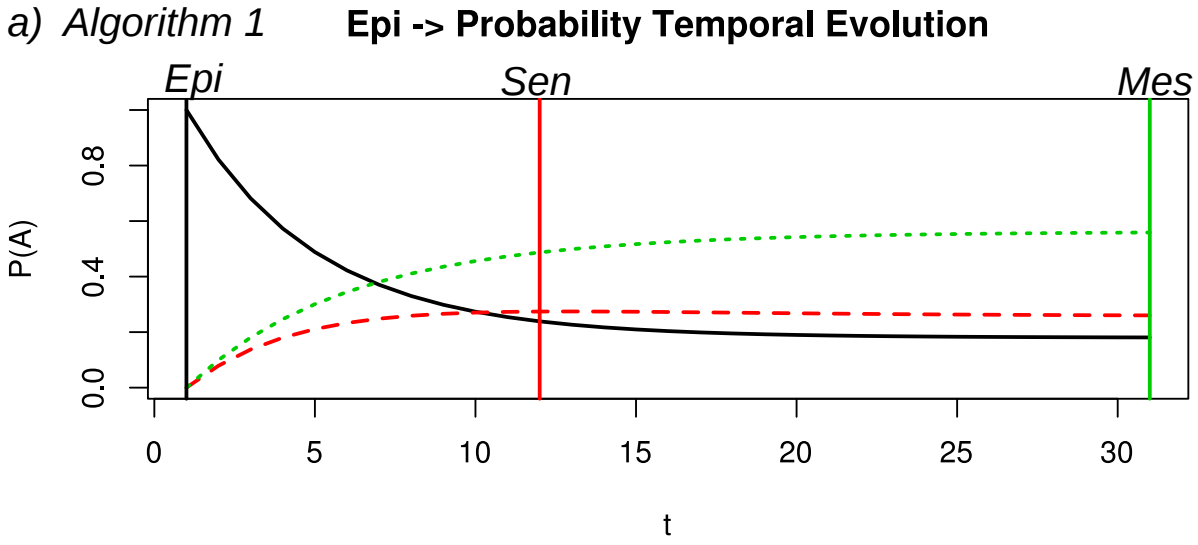

a)
state-space characterization
$x_1(t+1) = F_1(x_2)$
$x_2(t+1) = F_2(x_1)$
x_i=1 x_i=0
S1
S2
S3
S4
Aa
Ab
b)
inter-attractor transition (IAT)
inter-state transition (IAT)
Aa Ab
Aa raa rab
Ab rba rbb
S1 S2 S3 S4
S1 r11 ⋯ r14
S2 ⋮ ⋱ ⋮
S3
S4 r41 ⋯ r44
c)
attractor time-order
attractor global ordering
probabilistic landscape
da->b > 0
da->b > 0


a)
Estimation Algorithm 1
b)
Estimation Algorithm 2
c)
Estimation Algorithm 1
d)
Estimation Algorithm 2
e)
Estimation Algorithm 1
f)
Estimation Algorithm 2
dist(i,n)
0.00
0.10
0.20
0.30
0e+00
2e+05
4e+05
6e+05
8e+05
1e+06
0
200
400
600
800
1000
Estimate Nsteps(i)

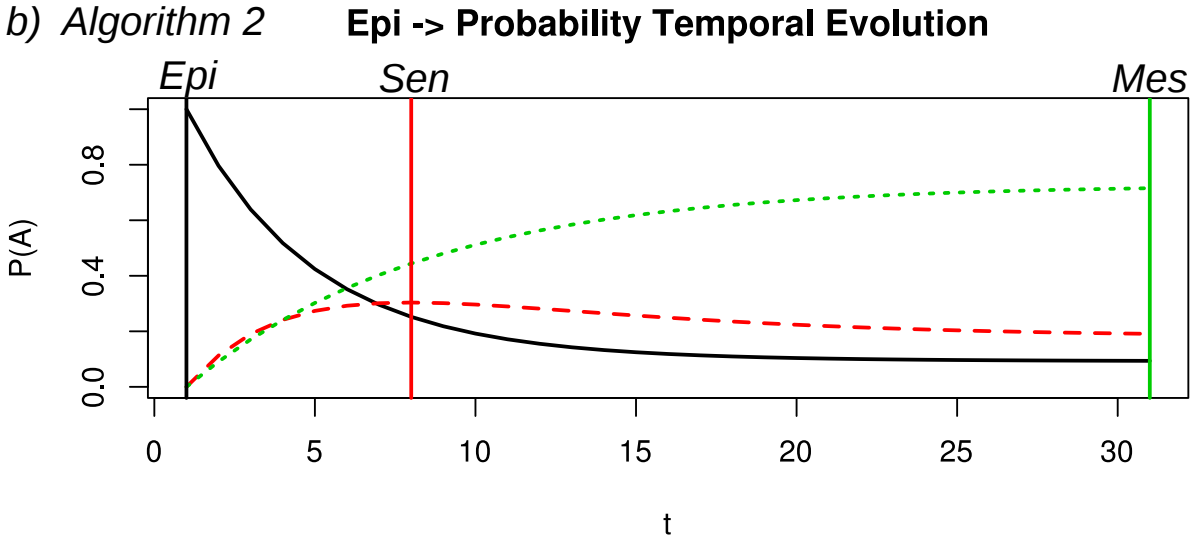

a)
Comparison Algorithm 1 and 2
dist(A1i,A2n)
0.00
0.10
0.20
0.30
0e+00
2e+05
4e+05
6e+05
8e+05
1e+06
A1 Estimate Nsteps(i)
b)
Comparison Algorithm 2 and 1
dist(A2i,A1n)
0.00
0.10
0.20
0.30
0
200
400
600
800
1000
A1 Estimate Nsteps(i)
c)
Comparison Algorithm 1 and 2
dist(A1i,A2n)
0.00
0.10
0.20
0.30
0e+00
2e+05
4e+05
6e+05
8e+05
1e+06
A1 Estimate Nsteps(i)
d)
Comparison Algorithm 2 and 1
dist(A2i,A1n)
0.00
0.10
0.20
0.30
0
200
400
600
800
1000
A1 Estimate Nsteps(i)
e)
Comparison Algorithm 1 and 2
dist(A1i,A2n)
0.00
0.10
0.20
0.30
0e+00
2e+05
4e+05
6e+05
8e+05
1e+06
A1 Estimate Nsteps(i)
f)
Comparison Algorithm 2 and 1
dist(A2i,A1n)
0.00
0.10
0.20
0.30
0
200
400
600
800
1000
A2 Estimate Nsteps(i)


TELasa
p53
NF-kB
ESE-2
Snai2
Cyclin
p16
Rb
E2F

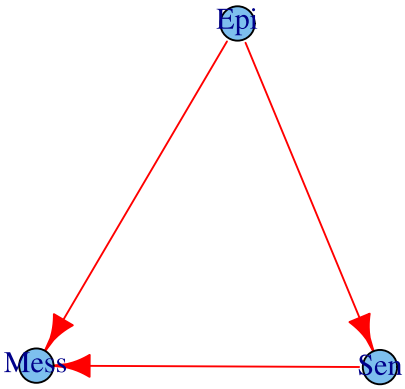
a) Algorithm 1



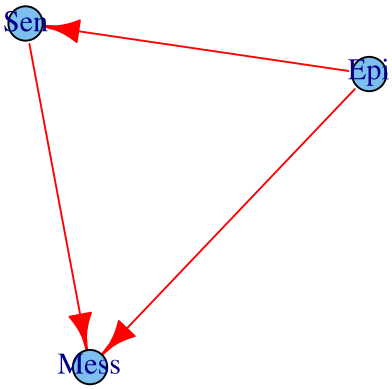
b) Algorithm 2

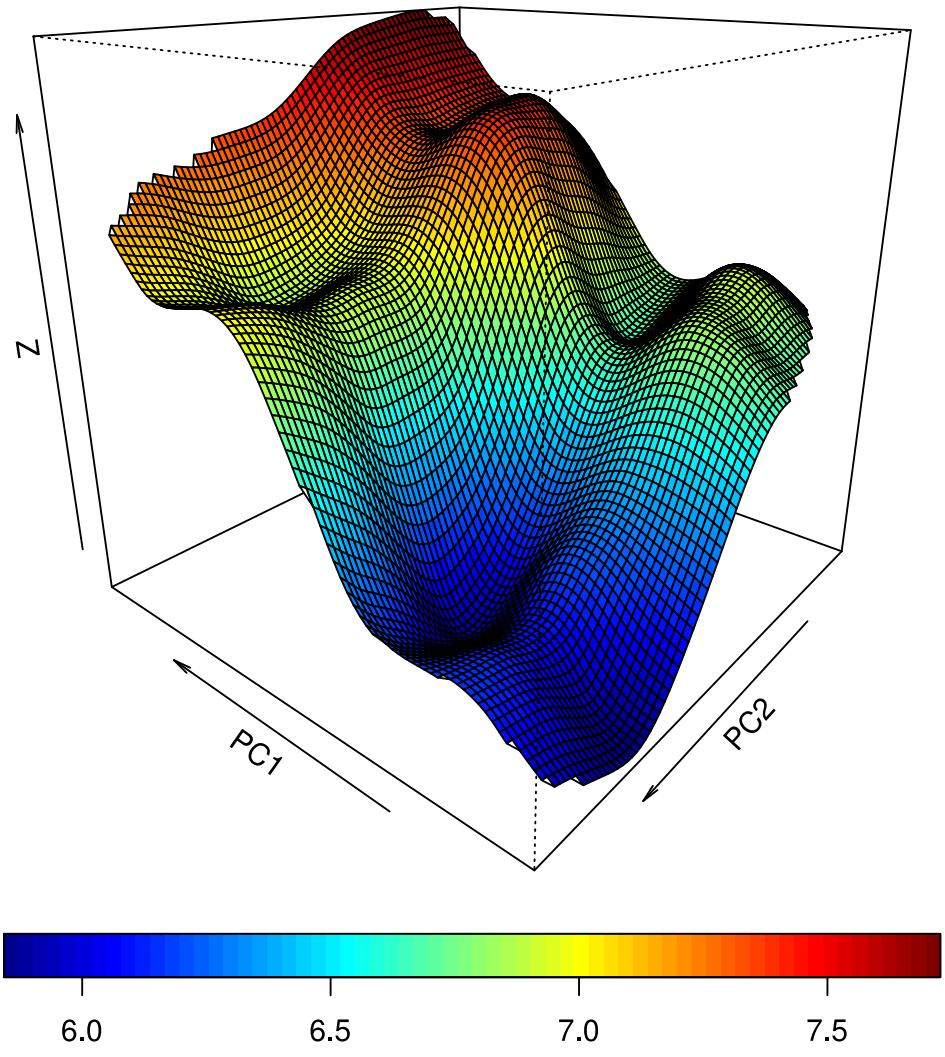


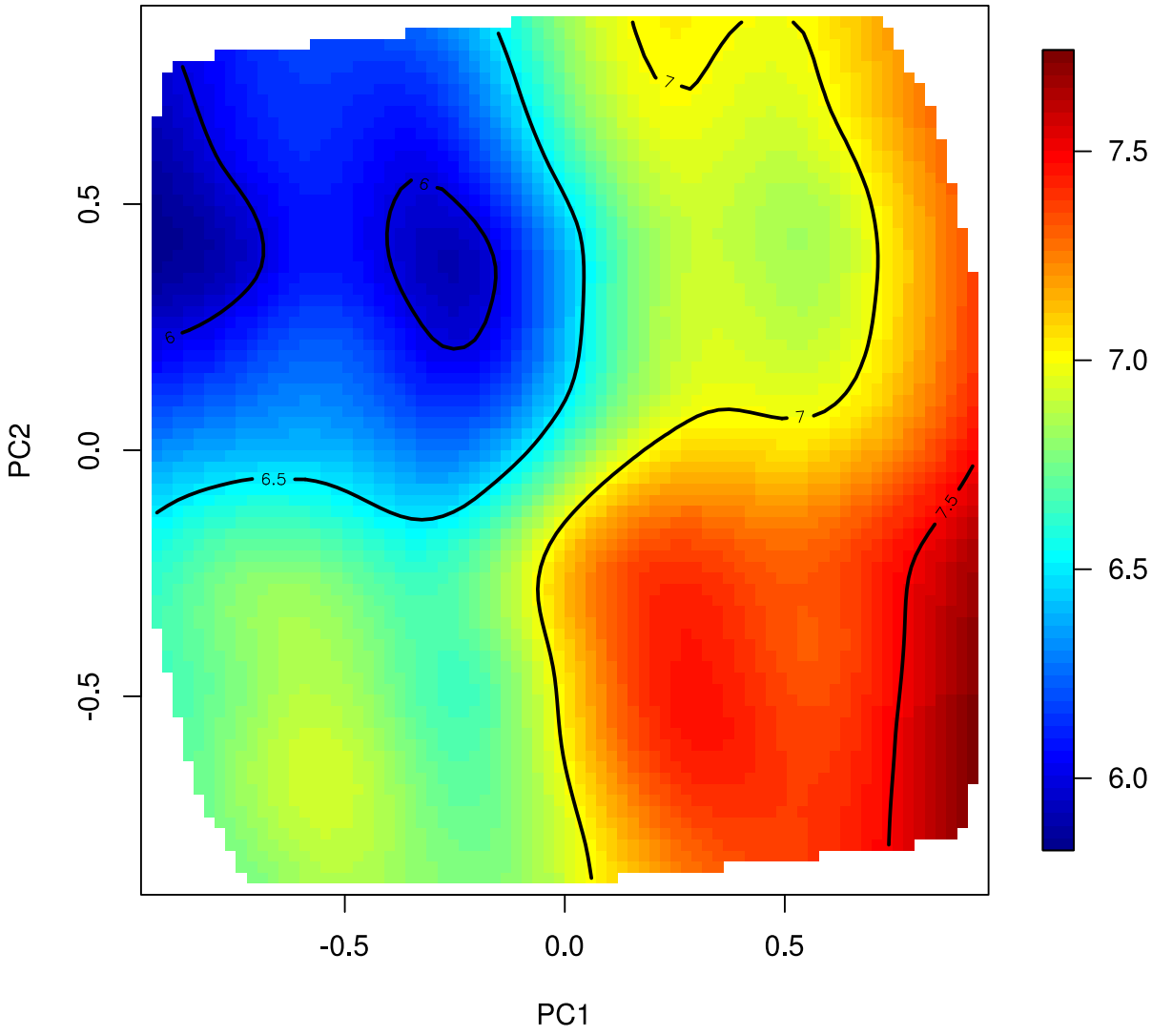
a) Algorithm 1



a) Algorithm 2




Z
PC1
PC2
6.0
6.5
7.0
7.5


PC2
0.5
0.0
-0.5
-0.5
0.0
0.5
PC1
6
6
6.5
7
7
7.5
7.5
7.0
6.5
6.0

MBE Advance Access published December 4, 2013

Molecular Evolution Constraints in the Floral Organ Specification Gene Regulatory Network Module across 18 Angiosperm Genomes

Jose Davila-Velderrain,[1,2] Andres Servin-Marquez,[3] and Elena R. Alvarez-Buylla*,[1,2]

[1]Instituto de Ecología, Universidad Nacional Autónoma de México, México, D.F., México

[2]Centro de Ciencias de la Complejidad, C3, Universidad Nacional Autónoma de México, México, D.F., México

[3]Facultad de Ciencias Biológicas, Universidad Autónoma de Nuevo León, San Nicolás de los Garza, Nuevo León, México

***Corresponding author:** E-mail: eabuylla@gmail.com.

Associate editor: Michael Purugganan

Abstract

The gene regulatory network of floral organ cell fate specification of *Arabidopsis thaliana* is a robust developmental regulatory module. Although such finding was proposed to explain the overall conservation of floral organ types and organization among angiosperms, it has not been confirmed that the network components are conserved at the molecular level among flowering plants. Using the genomic data that have accumulated, we address the conservation of the genes involved in this network and the forces that have shaped its evolution during the divergence of angiosperms. We recovered the network gene homologs for 18 species of flowering plants spanning nine families. We found that all the genes are highly conserved with no evidence of positive selection. We studied the sequence conservation features of the genes in the context of their known biological function and the strength of the purifying selection acting upon them in relation to their placement within the network. Our results suggest an association between protein length and sequence conservation, evolutionary rates, and functional category. On the other hand, we found no significant correlation between the strength of purifying selection and gene placement. Our results confirm that the studied robust developmental regulatory module has been subjected to strong functional constraints. However, unlike previous studies, our results do not support the notion that network topology plays a major role in constraining evolutionary rates. We speculate that the dynamical functional role of genes within the network and not just its connectivity could play an important role in constraining evolution.

Key words: **gene regulatory network, flower development, molecular evolution, functional constraint.**

Introduction

An outstanding goal in molecular evolution is to bridge the gap between the study of individual molecules and the study of systems on higher levels of biological organization. In modern evolutionary studies, the limitations of considering genes as individual entities upon which evolutionary forces act independently are becoming generally accepted. The emerging picture is that in which evolutionary forces, functional constraints, and molecular interactions are conditionally dependent on the systems level (Cork and Purugganan 2004). Following this line of research, several studies have analyzed molecular evolution at the pathway or network level (see, e.g., Hahn et al. 2004; Alvarez-Ponce et al. 2009; Jovelin and Phillips 2009; Yang et al. 2009; Montanucci et al. 2011; Alvarez-Ponce 2012). Most studies support the idea that evolutionary forces acting on genes are in close relation with the structure/topology of their functional network.

Previous network-based molecular evolutionary studies have focus on investigating networks in relation to the evolutionary rates of their genes based on large-scale molecular networks (Fraser et al. 2002; Agrafioti et al. 2005; Hahn and Kern 2005; Lemos et al. 2005; Alvarez-Ponce and Fares 2012). Recently, similar analysis have been applied to well-characterized, relatively small pathways (Alvarez-Ponce et al. 2009, 2011; Casals et al. 2011; Fitzpatrick and O'Halloran 2012; Lavagnino et al. 2012; Invergo et al. 2013). Both approaches have uncovered interesting yet preliminary patterns (see Montanucci et al. 2011 and references therein). The conclusion, so far, appears to be that evolutionary pressures acting on genes are in close relation with the structure of their functional network. But contrasting results have been found in several cases, and when considering the latter, there is no general consensus for the relationship between network properties and the molecular evolution of its components: different patterns have been found for different interacting systems and different species sets. Thus, the need for resolution of contrasting results and the search of robust evolutionary patterns call for new studies. It has been suggested that the analysis of new pathways might help to uncover general patterns and to disentangle topological restrictions of networks from the biological properties and functions (Montanucci et al. 2011). Here, we argue that the study of the molecular evolution of the genes involved in regulatory modules that have been uncovered with dynamical gene regulatory network (GRN) models could help uncover general evolutionary principles, given that such models allow a

Article

rigorous distinction between structure and function. In contrast to schematic representations that depict gene regulatory interactions, dynamic models may consider the nonlinear aspects of regulation and explore the way gene expression changes in time, both in wild-type and perturbed simulated systems (Alvarez-Buylla et al. 2010). Nevertheless, to the best of our knowledge, a network-based molecular evolutionary study is lacking for the case of experimentally grounded and functionally validated dynamic GRN models.

It is generally accepted that GRNs are underlying molecular systems orchestrating developmental processes (Huang and Kauffman 2009; Alvarez-Buylla et al. 2010). On the other hand, it has been suggested that the specific nature of evolutionary forces acting on the component genes depends largely on the function of the interacting system (Cork and Purugganan 2004). In this work, we follow a similar approach to that of previous network-level evolutionary studies; but instead of analyzing a new metabolic pathway, we focus on the molecular evolution and network properties of a well-studied GRN module: the experimentally grounded floral organ cell fate specification determination GRN (FOS-GRN) (see Espinosa-Soto et al. 2004; Alvarez-Buylla et al. 2010 for updates).

The FOS-GRN (fig. 1) integrates molecular genetic data for the ABC genes and their main interactors in *A. thaliana*. This GRN includes key regulators underlying the transition from the shoot apical meristem once it produces the apical inflorescence meristem with the flower primordia in its flanks (flowering locus t [*FT*], terminal flower1 [*TFL*], embryonic flower1 [*EMF1*], LEAFY [*LFY*], APETALA1 [*AP1*], fruitfull [*FUL*]), the ABCs and some of their interacting genes (APETALA1 [*AP1*], APETALA3 [*AP3*], PISTILLATA [*PI*], APETALA2 [*AP2*], AGAMOUS [*AG*], SEPALLATA [*SEP*]), as well as some genes that link floral organ specification to other modules regulating primordia formation and homeostasis (*AG* and *WUS*) and to some regulators of organ boundaries (*UFO*). From the 15 genes, 6 are members of the MADS-box protein family (*AG*, *AP1*, *AP3*, *PI*, *SEP*, *FUL*) and belong to five different subfamilies (AG, SQUA, GLO, DEF, and AGL2) within the clades of MADS-box genes (Becker and Theissen 2003).

The model was proposed on the basis of experimental data for these 15 genes in the model plant *A. thaliana*. Among the 15 genes, 5 are grouped into three classes (A-type, B-type, and C-type) whose combinations, described by the ABC model, are necessary for floral organ cell specification (Coen and Meyerowitz 1991). A-type genes (*AP1* and *AP2*) are necessary for sepal specification, A-type together with B-type (*AP3* and *PI*) for petal specification, B-type and C-type (*AGAMOUS*) for stamen specification, and the C-type gene (*AG*) alone for carpel primordia cell specification. Although the ABC model of flower development was published more than 20 years ago, it was just recently that the model of the FOS-GRN provided a sufficient explanation for the observed ABC patterns and the stable gene expression configurations observed during early flower development in *Arabidopsis* (Mendoza and Alvarez-Buylla 1998; Espinosa-Soto et al. 2004; and updates and review in Alvarez-Buylla et al. 2010). The network

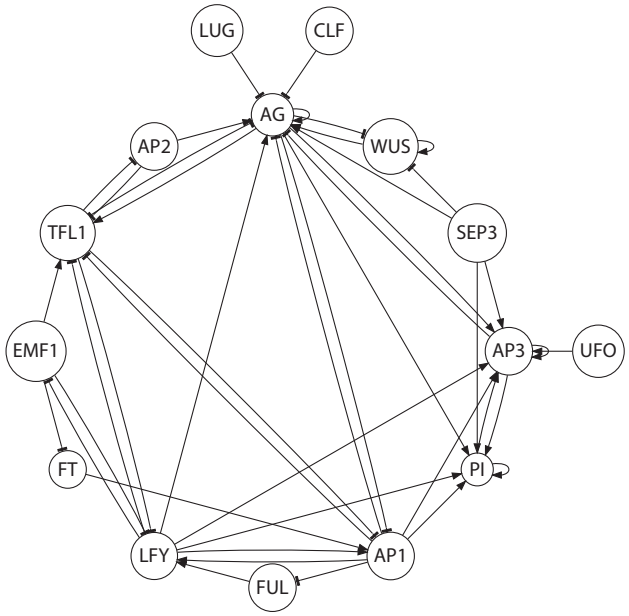


Fig. 1. Graph representation of the FOS-GRN. Arrows and blunt-ended edges correspond to activating and repressing interactions, respectively.

has been studied from different perspectives (Alvarez-Buylla et al. 2008; Sanchez-Corrales et al. 2010; Villarreal et al. 2012), and the results of multiple studies have shown that its dynamical behavior is robust enough as to predict the observed phenotypes both in wild-type and several mutant conditions. In other words, there is enough evidence to sustain the claim that the 15 genes involved in the network form a core regulatory module responsible for primordial cell fate determination during early stages of flower development. We reasoned that such a functional constraint could play a strong role in constraining evolutionary rates at the molecular level. Based on this idea, here we addressed whether orthologous genes of the FOS-GRN were found and conserved in distantly related angiosperm species, and then we addressed the evolutionary forces that could have shaped its evolution under the hypothesis that positive Darwinian selection would not be a prevailing force.

A large number of the genes involved in floral development belong to the eukaryotic MADS-box gene family (Riechmann et al. 1997). Most studies on the molecular basis of floral development focus on these genes, particularly floral homeotic genes such as *AGAMOUS* (*AG*), *APETALA3* (*AP3*), *PISTILLATA* (*PI*), and several *AGAMOUS*-like genes (Lawton-Rauh et al. 2000). Background information on genetic and expression analyses indicate that members of a floral homeotic gene group tend to share similar developmental functions in flower and inflorescence morphogenesis (Purugganan et al. 1995; Purugganan 1997), thus reflecting high conservation among evolutionarily related regulatory genes. Previous studies on the evolutionary forces acting on some of the genes involved in flower development have focused on intraspecific population genetics data (Purugganan and Suddith 1999) or data from two closely related species (Yang et al. 2011). These studies have shown that although most floral genes have evolved under strong purifying

selection, some show elevated nonsynonymous substitution rates and/or positively selected sites. However, given that these molecular evolutionary studies have focused mostly on closely related species, it is not known whether the complete set of genes conforming the FOS-GRN are globally conserved among flowering plants. In order to first explore this possibility, here we follow a comparative genomics approach, and, unlike previous work, we study the molecular evolution of the network over a broad taxonomic distance involving monocots and dicots; the recent completion, annotation, and analysis of the genomes of several flowering plant species has provided the opportunity to do so.

In summary, the aim of this work was 3-fold: 1) to explore the degree of conservation of the genes involved in the FOS-GRN, 2) to uncover the prevailing molecular evolutionary forces acting upon its genes, and 3) to study the evolutionary constraints that its network properties and known biological function impose to the molecular evolution of its components. With this in mind, we first searched for the homologs of the genes in the *A. thaliana* FOS-GRN in all the flowering species with a sequenced and annotated genome available (a total of 18; see fig. 2 for the species used and their placement in angiosperm phylogeny). With the sequence data for the FOS-GRN genes, we measured the action of selective pressures on individual protein-coding genes through the estimation of synonymous and nonsynonymous substitution rates (dS and dN, respectively) when comparing among species. The ratio dN/dS measures the strength and nature of the evolutionary forces indicating positive selection, neutral evolution, or purifying selection when it is higher, equal, or lower than 1, respectively. Both an overall ratio for the entire coding sequence of a gene and estimates considering variation of the ratio among sites were calculated (Yang and Bielawski 2000). We then calculated molecular conservation features other than evolutionary rates for each gene and asked whether these features in addition to the evolutionary parameters (dN, dS, dN/dS) show a pattern of association with the known biological functions of the genes. Finally, we addressed whether the forces that have shaped the evolution of the genes during the divergence of angiosperms were correlated to the placement of each gene within the FOS-GRN.

Results

Identification of the FOS-GRN Genes in Flowering Plant Genomes

The experimentally grounded FOS-GRN proposed by Espinosa-Soto et al. was used as a reference (Espinosa-Soto et al. 2004; and updated in Alvarez-Buylla et al. 2010). The original network proposed for *A. thaliana* has 15 genes and their regulatory (activating or inhibitory) interactions (supplementary table S1, Supplementary Material online). In order to study the conservation of the genes in the network across species, we conducted homology analysis using the Plaza Comparative Genomics Platform (Proost et al. 2009) (see Materials and Methods). For each gene in the network, a total of 418 putative homologs (orthologs and in-paralogs) of the 15 *A. thaliana* (Ath) FOS-GRN genes were identified in the genomes of the other 17 flowering plant species: *Arabidopsis lyrata* (Aly), *Brachypodium distachyon* (Bdi), *Carica papaya* (Cpa), *Fragaria vesca* (Fve), *Glycine max* (Gma), *Lotus japonicus* (Lja), *Malus domestica* (Mdo), *Manihot esculenta* (Mes), *Medicago truncatula* (Mtr), *Oryza sativa japonica* (Osj), *Oryza sativa indica* (Osi), *Populus trichocarpa* (Ptr), *Ricinus communis* (Rco), *Sorghum bicolor* (Sbi), *Theobroma cacao* (Tca), *Vitis vinifera* (Vvi), and *Zea mays* (Zma) (see fig. 2). These results correspond to the preliminary network conservation data and were organized in the form of a conservation matrix (also called phylogenetic profile) where each row represents a gene vector composed by a set of characters {0, 1, 2, 3, 4} representing the absence (0), presence (1), or the total number of in-paralogs (2, 3, 4) of each gene; and each column represents a species (supplementary table S2, Supplementary Material online). All FOS-GRN genes studied, with the exception of *EMF1*, have orthologs in all 18 genomes. The gene *EMF1* was not found as an ortholog of the *EMF1* gene in *A. thaliana* (AT5G11530) among the monocot plants: *B. distachyon*, *O. sativa japonica*, *O. sativa indica*, *S. bicolor*, and *Z. mays*. However, following the same methodology, but using instead the corresponding protein sequence of the gene *EMF1* reported for *O. sativa* (OS01G12890) as query, putative orthologs were found in all four cases. For the only case of this gene (*EMF1*), it was discovered that there exists one orthologous group for dicots and a different group for monocots. The relationship between both groups is not clear and will be studied in subsequent studies.

Manual Curation of Putative In-paralogs

The preliminary conservation data of the proteins in the FOS-GRN of *A. thaliana* were manually curated to produce the final conservation data of the proteins in the FOS-GRN reported here in the form of a conservation matrix (fig. 2) and the corresponding list of gene IDs (supplementary table S4, Supplementary Material online). Certain proteins were eliminated from the list due to evidence of partial gene copies or annotation errors (see Materials and Methods). We found that all the FOS-GRN genes have homologs in all 18 genomes searched. Results also show that all the genes underwent a number of duplication and/or loss events. The detailed evolutionary processes (e.g., duplication, loss, and pseudogenization) leading to the expansion of the network across angiosperms will be explored in a future study.

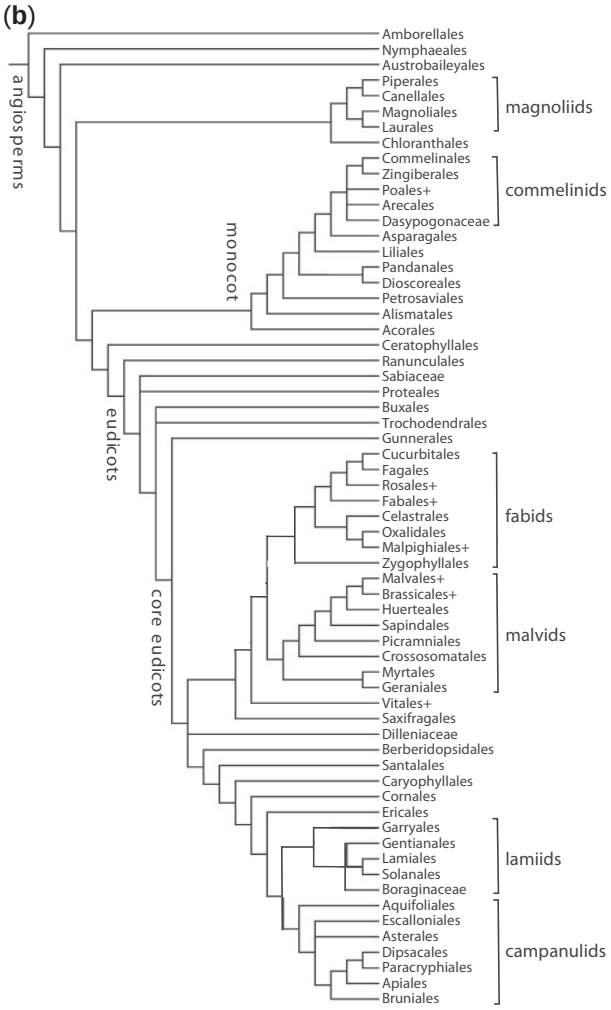
Molecular Evolutionary Analysis of the FOS-GRN

The nonsynonymous (dN) to synonymous (dS) substitution rate ratio (dN/dS) was calculated in order to infer the impact of natural selection on the FOS-GRN. The values of the overall ratio dN/dS range from 0.05936 for *PI* to 0.39577 for *EMF1*, suggesting that purifying selection or selection constraint best explains the evolution of the genes in the FOS-GRN (table 1). Given that the estimation of an overall dN/dS for the whole coding sequence is a very conservative measure of positive selection (Yang and Bielawski 2000), estimates that account for variation in dN/dS among sites in order to detect specific sites that could have been fixed by positive selection were also

(a)

Gene	Aly	Ath	Bdi	Fve	Gma	Lja	Mdo	Mes	Mtr	Osj	Osi	Ptr	Rco	Sbi	Tca	Vvi
AG	1	1	2	1	2	2	2	1	1	2	1	1	1	2	1	1
AP1	1	1	2	1	2	1	2	2	2	2	1	2	1	2	1	1
AP2	1	1	1	1	4	1	4	2	1	1	1	2	2	1	2	2
AP3	1	1	1	1	2	1	1	2	1	1	1	1	1	1	1	1
CLF	1	1	1	1	2	1	2	1	1	1	1	2	1	1	1	1
EMF1	1	1	1*	1	1	1	3	1	2	1*	1*	2	1	1*	1	1
FT	1	1	2	1	2	1	2	1	1	3	2	3	4	2	1	1
FUL	1	1	1	1	3	1	2	2	1	1	1	1	1	1	1	1
LFY	1	1	1	3	2	1	2	2	1	1	1	1	1	1	1	1
LUG	1	1	3	2	4	1	1	1	2	1	2	2	1	2	2	1
PI	1	1	2	1	4	1	1	2	1	2	1	2	1	2	1	1
SEP	1	1	1	1	2	1	2	1	2	1	1	2	1	1	1	1
TFL1	1	1	1	1	2	1	2	3	1	2	2	2	1	2	2	1
UFO	1	1	1	2	2	1	2	1	1	1	1	2	1	1	1	1
WUS	1	1	1	3	2	2	5	3	2	1	1	4	1	1	2	3

(b)



(c)

Species	APG III
Carica papaya (Cpa)	Brassicales
Arabidops is thaliana (Ath)	Brassicales
Arabidops is lyrata (Aly)	Brassicales
Manihot es culenta (Mes)	Malpighiales
Glycine max (Gma)	Fabales
Lotus japonicus (Lja)	Fabales
Medicago truncatula (Mtr)	Fabales
Populus trichocarpa (Ptr)	Malpighiales
Ricinus communis (Rco)	Malpighiales
Theobroma cacao (Tca)	Malvales
Brachypodium distachyon (Bdi)	Poales
Oryza sativa japonica (Osj)	Poales
Oryza sativa indica (Osj)	Poales
Sorghum bicolor (Sbi)	Poales
Zeamays (Zma)	Poales
Fragaria vesca (Fve)	Rosales
Malus domestica (Mdo)	Rosales
Vitis vinifera (Vvi)	Vitales

Fig. 2. Gene conservation data, species used, and their placement in Angiosperm phylogeny. (*a*) Conservation matrix of the genes involved in the FOS-GRN across Angiosperm species (*Genes were identified using *Oryza sativa* EMF1 protein (OS01G12890) for homology search; $^{+}$Families considered in the analysis). (*b*) Angiosperms phylogeny APG III according to Bremer et al. (2009). (*c*) Species used in the analysis.

calculated. Results showed that the genes *UFO*, *FT*, and *CLF* yielded a marginal significant *P* value when comparing the model M8 assuming positive selection with the null model M7 of the program CODEML (see Materials and Methods). However, the test was no longer significant after correcting for multiple comparisons. For all 15 genes, the models M2a was not significantly better than the null model M1a (supplementary table S5, Supplementary Material online). The overall dN/dS, dN, and dS were computed for each gene under the M0 model (table 1). The genes of the FOS-GRN are subject to strong purifying selection with an overall mean dN/dS of 0.124. Overall dN/dS values are plotted in figure 3; from the 15 genes, 13 (86.66%) have a dN/dS value <0.15.

Table 1. Evolutionary Parameters of the FOS-GRN Genes.

Gene	Locus	Protein Length	Percent of Analized Codons	dN	dS	dN/dS
AP1	AT1G69120	256	89	0.7683	6.1525	0.12487
AP2	AT4G36920	432	80	0.6095	5.6578	0.10773
AP3	AT3G54340	232	93	0.7713	8.0723	0.09555
CLF	AT2G23380	902	67	0.6369	5.386	0.11824
EMF1	AT5G11530	1096	66	3.8105	9.6281	0.39577
FT	AT1G65480	175	99	0.4509	5.9891	0.07529
FUL	AT5G60910	242	95	0.8261	7.0456	0.11725
LFY	AT5G61850	420	83	0.715	8.5201	0.08392
LUG	AT4G32551	931	78	0.5995	5.2033	0.11522
PI	AT5G20240	208	58	0.602	10.1413	0.05936
SEP	AT1G24260	250	90	0.6172	7.9816	0.07733
TFL1	AT5G03840	177	97	0.5	5.973	0.0837
UFO	AT1G30950	442	84	0.9109	8.4945	0.10723
WUS	AT2G17950	292	51	2.615	12.799	0.20431

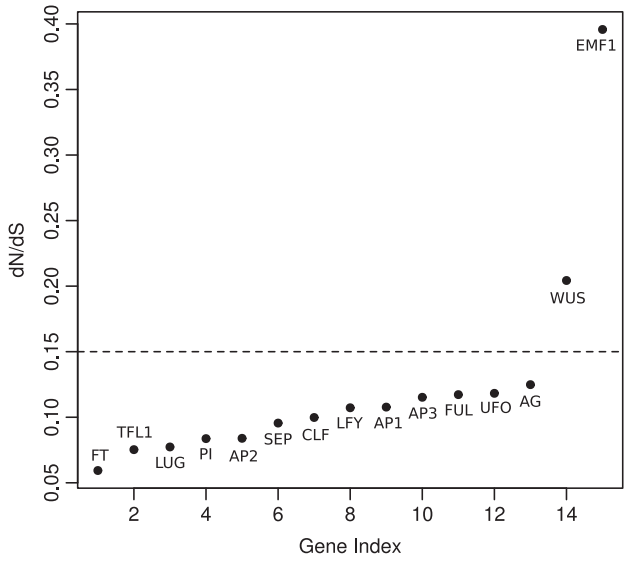


Fig. 3. Calculated dN/dS values sorted in increasing order. The horizontal dotted line is plotted to show that, from the 15 genes, 13 (86.66%) have a dN/dS value <0.15. Plotted values were calculated using the M0 model.

Analysis of the Classes of Genes

To test whether the measures of dN/dS, dN, or dS were statistically different between two gene classes, the ABC genes and the additional genes in the network, a Kruskal–Wallis test was performed. Although the genes *EMF1* and *WUS* showed higher dN/dS values than the 86.66% of the genes, the test gave no significant differences in dN/dS, dN, or dS between the classes. Means and *P* values are shown in supplementary tables S6 and S7, Supplementary Material online.

Model-Based Clustering of Sequence Conservation Features

During initial exploratory data analysis, it was observed that protein and DNA coding sequences of some of the genes in the FOS-GRN across the angiosperm species show interesting patterns in measures of conservation other than the evolutionary parameters (supplementary figs. S1–S7, Supplementary Material online). The following conservation features were calculated (see Materials and Methods): the degree of variability in protein size of each protein over all species (measured by the coefficient of variation), mean protein pairwise sequence distances, mean protein sequence distance, and mean DNA sequence distance (table 2). Given such data, the following question raised: is there an association between such conservation patterns and the functional classification of the proteins in the network?

In order to explore this possibility, a model-based clustering analysis was applied. Clustering is the process of grouping similar objects together. Here, a feature-based clustering approach was used, in which an $N \times D$ feature matrix is used as input (Murphy 2012). A feature matrix was assembled where each of the N rows represents a particular gene and the D columns corresponded to the conservation features listed above, together with an additional column corresponding to the dN/dS data (table 2). In other words, each row represents a conservation feature vector for each gene. This analysis does not make any assumption about the prior known functional category of the genes. Instead, it divides the genes into clusters according to the similarity among their feature vectors. The analysis was restricted to include all but the *EMF1* and *WUS* genes: when all the genes were included, an additional cluster was invariably obtained for each of the two genes (*EMF1* and *WUS*) given their high dN/dS and interspecies sequence distances (data not shown). Interestingly, the methodology uncovered three clusters (fig. 4): one corresponding the genes *AG*, *AP1*, *AP2*, and *PI* (circles); one for the genes *FUL*, *LFY*, *UFO*, and *AP3* (triangles); and the last one to the additional genes *CLF*, *FT*, *LUG*, *SEP*, and *TFL1* (squares). The four genes in the first cluster correspond to ABC floral organ identity genes. While the genes in the second cluster, except *AP3*, floral meristem identity genes (Krizek and Fletcher 2005). These results suggest an association between molecular size and sequence conservation features, evolutionary rates, and functional category. Those genes with a well-characterized function

Table 2. Gene Conservation Features.

Gene	No. Protein Sequences	Protein Size (CV)	Protein Mean Distance	DNA Codon Mean Distance	dN/dS
AG	30	0.241130802	0.437267	0.4316755	0.09981
AP1	30	0.210350092	0.4575694	0.4514316	0.12487
AP2	30	0.095096262	0.4422398	0.4184613	0.10773
AP3	22	0.099160315	0.4942555	0.4455772	0.09555
CLF	22	0.354842355	0.4084511	0.3709341	0.11824
EMF1	28	0.323552388	1.742475	1.2504943	0.39577
FT	31	0.254844679	0.2390374	0.3548233	0.07529
FUL	22	0.18187526	0.464033	0.4358368	0.11725
LFY	25	0.182826509	0.4513709	0.4390543	0.08392
LUG	37	0.236524789	0.3276286	0.3408934	0.11522
PI	29	0.184321532	0.4537865	0.3875463	0.05936
SEP	25	0.19888967	0.3333186	0.3837757	0.07733
TFL1	27	0.009807079	0.2475236	0.324812	0.0837
UFO	23	0.037201769	0.6089123	0.5777019	0.10723
WUS	36	0.17482874	1.1723443	0.8080616	0.20431

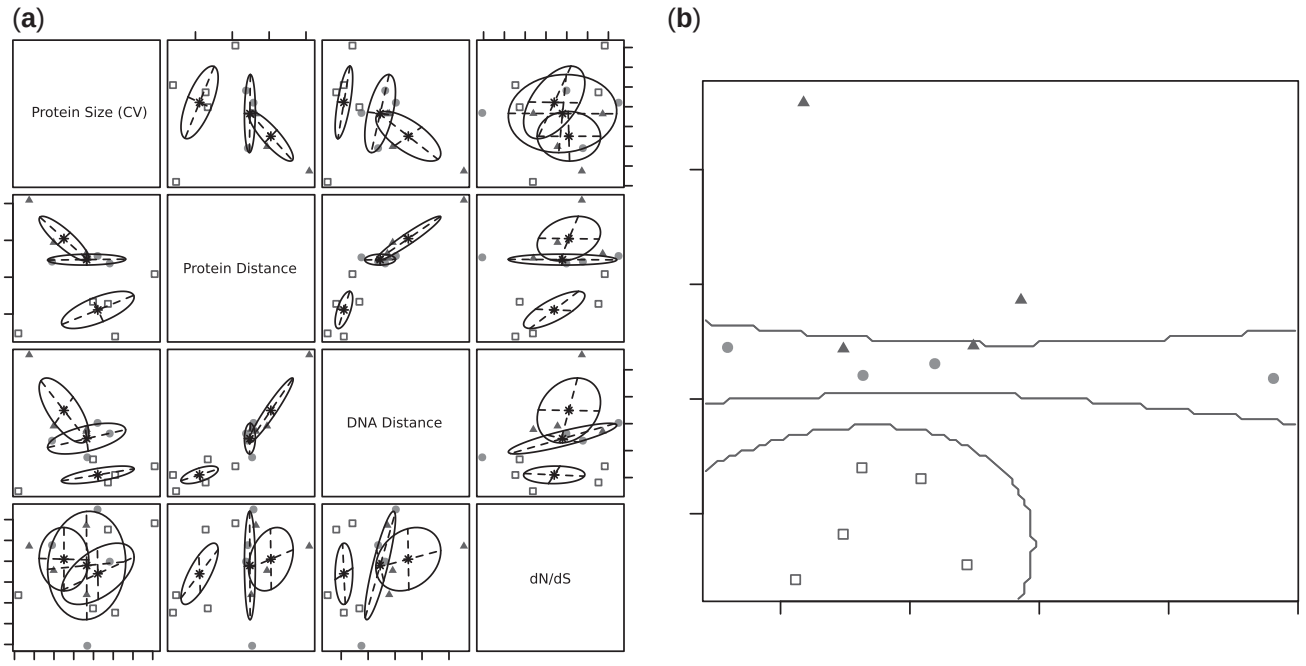


Fig. 4. Output from the model-based clustering analysis. (*a*) Scatterplot matrix for conservation features with points (genes) marked according to the corresponding cluster; the ellipses shown are the multivariate analogs of the standard deviations for each mixture component. (*b*) Data projection on a dimension reduced subspace. Clustering structure and boundaries are shown; genes are marked according to the corresponding cluster.

(e.g., a direct involvement in the processes of floral or meristem identity) share more similar conservation features among them than with the additional interacting genes which are known to be involved in various processes. Genes in the last cluster are known to integrate the flowering process with upstream signaling mechanisms and either promote (e.g., *FT*) or inhibit (*TFL1*) flower organ development. Figure 4*b* shows a two-dimensional projection of the feature vector along with the corresponding classification boundaries; it is interesting to note that the boundaries between the meristem (triangles) and flower identity (circles) clusters merge and are clearly separated from the third cluster (squares). This is consistent with the known biological mechanisms where genes such as *AP1* participate as both meristem and floral organ identity genes.

Given that clustering is an unsupervised learning technique, it is hard to evaluate the quality of the output on any given method. One way to do so is to rely on some external form of data with which to validate the method. In the case in point, labels representing functional categories can be assigned to each gene. Each gene was labeled with one of the three categories: floral organ identity, floral meristem identity, and other. The clustering was then compared with the labels using a standard metric: the Rand index (see Materials and Methods). This metric was calculated for the output of the clustering. Then, its statistical significance was

assessed through their frequency sampling distribution computed using a bootstrap resampling method (Murphy 2012). The observed clustering decisions are highly significant (P value = 0.0002). Thus, there is statistical support for an association between the molecular conservation features, the evolutionary rates, and the functional category of the genes in the FOS-GRN.

The Strength of the Purifying Selection and Network Structure

Each node in the network was characterized by a set of features including the molecular evolutionary parameters (dN, dS, and dN/dS) and its placement within the network topology, using measures such as centrality, degree, closeness, betweenness, and eccentricity (see Materials and Methods). GRNs contain directed interactions with either an inhibitory or an activating character. Given that the dynamical behavior of GRNs is associated with the type of interactions within the network, the topological network properties, out-degree, in-degree, activating in-degree, and inhibitory in-degree, were also included as features (supplementary table S8, Supplementary Material online). Once the evolutionary parameters and the network topological features were calculated, the goal was to answer the following questions: 1) Is there a relationship between the evolutionary parameters and the network nodes topological location within the FOS-GRN? 2) How strong is the relationship found, if any? 3) Which network topological features contribute the most to evolutionary rates?

A relationship between each of the evolutionary parameters and each of the node's topological features within the FOS-GRN was tested. Assuming an approximately linear relationship, model coefficients were estimated independently for each of the networks' topological features as single predictor variables of the evolutionary parameters. Hypothesis tests on the coefficients were performed in order to test whether or not there is a relationship between the variables in each case. Mathematically, this corresponds to testing whether the corresponding coefficient is equal to 0 or not. Details of the least squares models for the regression of dN/dS on each of the topological features used are provided in supplementary table S10, Supplementary Material online. Interestingly, the null hypothesis that the coefficient is equal to 0 could not be rejected for any case; consequently, a relationship between the dN/dS and any of the networks topological features tested could not be declared to exist, given the available data. The same analysis was applied individually to dN and dS as response variables. Only a marginal significant relationship (P value ~0.05) was found between dS and closeness. In a preliminary analysis, Spearman's rank correlation coefficients between the evolutionary parameters and the topological network properties were also calculated and are reported in supplementary table S9, Supplementary Material online. No significant correlation was found between the measures of centrality and the evolutionary estimates.

Similarity in Evolutionary Parameters of Interacting Genes

It has been suggested that interacting elements within a network share more similar values of evolutionary parameters within themselves than with noninteracting components (Alvarez-Ponce et al. 2009). In order to test whether this pattern is present in the FOS-GRN, two different approaches were applied: 1) the average absolute difference (AAD) of the value of the evolutionary parameters between interacting components in the networks was used as an statistic and compared with its null distribution in an ensemble of similar but random networks (Alvarez-Ponce et al. 2009), and 2) a matrix of pairwise shortest path distances between the genes in the network was compared with the matrices of pairwise absolute differences in evolutionary parameters (Montanucci et al. 2011). Using the former approach, an AAD of dN/dS of 0.0567 was calculated for the FOS-GRN. The histogram of the corresponding statistic on an ensemble of 100,000 random networks with the same number of nodes and interactions is shown in supplementary figure S8, Supplementary Material online. The simulated data follow closely a Gaussian distribution. The obtained data were used to estimate the probability of observing such a small value. Two approaches were followed: 1) calculating the fraction of random networks showing an AAD value $\leq$0.0567 and 2) calculating the probability of such a value using a Gaussian density function with an empirically estimated mean and standard deviation (supplementary fig. S8, Supplementary Material online). The resulting probabilities were 0.12768 and 0.12852, respectively.

For the second approach, a Mantel test comparing a matrix of pairwise distances between genes in the network and matrices of pairwise absolute differences in evolutionary parameters was applied for dN/dS, dN, and dS. The test found no significant correlation between distance and difference in any evolutionary parameter (supplementary table S11, Supplementary Material online). The results of both approaches do not support the hypothesis that neighboring genes share similar evolutionary constraints in the case of the FOS-GRN.

Discussion

The question of whether the role of regulators involved in the control of floral initiation is conserved across flowering plants has been raised recently in the literature (Wellmer and Riechmann 2010). Of particular interest is the situation of grass-like plants and other monocots, which are distantly related to *A. thaliana* and its relatives. Based on the identification of homologs of the main regulators involved in the control of floral initiation of *A. thaliana* in monocots as well as observations of expression patterns in different species, it has been suggested that many aspects of the topology of the floral transition network seem to be conserved between dicots and monocots (Wellmer and Riechmann 2010). However, empirical gene conservation data based on whole-genome analysis were lacking. Given the availability of multiple genomes of angiosperms—both monocots and dicots—a comparative genomics approach was possible and

enabled us to uncover a clearer picture of the conservation status of the regulators known to be involved in the control of floral initiation and floral organ specification in *A. thaliana* across angiosperms. We focused specifically on the regulators participating in the FOS-GRN model (see Espinosa-Soto et al. 2004; Alvarez-Buylla et al. 2010 for updates). Our results show that all the FOS-GRN genes have representatives in the 18 angiosperm species used in this study. The existence of all the genes in all the surveyed species, together with the high selective constraint level found in this study (mean dN/dS = 0.124), suggests that the FOS-GRN is functionally constrained across all these species belonging to nine families, nine orders, and both monocot and dicot species. This is consistent with what we might expect given the robustness of the FOS-GRN as a developmental regulatory module and the observed expression patterns of some of the genes of this GRN documented for different species (Espinosa-Soto et al. 2004; Alvarez-Buylla et al 2010). These results, however, do not provide information of whether or not there are considerable differences in network circuitry among species. The empirical data obtained here may serve, nonetheless, as a basis to explore the dynamical behavior of the corresponding FOS-GRN in different species under the assumption of conserved interactions among network components. Indeed, further model refinements as well as phenotypic validations and testable predictions could be generated following such a theoretical approach.

Our results also show that the genes in the FOS-GRN have undergone a number of duplication and/or loss events. The evolutionary history of MADS-box genes involved in flowering has been extensively studied with phylogenetic approaches (see, e.g., Alvarez-Buylla et al. 2000). A complex history of gene duplications within the AP1/FUL clade during angiosperm evolution is well documented (Preston and Kellogg 2007). The results of gene conservation obtained in this work suggest a similar complex history for most of the genes of the other gene families in the FOS-GRN. Furthermore, it is well known that some of the species included in the study have shared whole-genome duplication (WGD) events. For example, *A. thaliana* has experienced at least three WGD events—two recent events since its divergence from other members of the Brassicales clade and a more ancient event shared with most, if not all, eudicots (Bowers et al. 2003). A WGD event occurred more than once before the split between *A. thaliana* and *A. arenosa* (Ha et al. 2009); consequently, the two *Arabidopsis* species included in the analysis have shared WGD events which are not shared with the other species. This evolutionary scenario may partially account for the complex pattern of duplications observed in the conservation data; unfortunately, it also makes it difficult to establish clear relationships of orthology. The empirical conservation data reported herein thus serve as a basis for further phylogenetic studies which are needed in order to better explain the processes leading to the conservation and expansion of the FOS-GRN across angiosperms. The data concerning the overall conservation of the FOS-GRN genes obtained here suggest interesting questions for future investigation in diverse angiosperm species, such as addressing whether the interactions of the flower organ identity genes and their interacting partners are conserved among monocots and dicots or not. What is the role of the duplicated genes in the dynamics of the FOS-GRN? Does such gene redundancy increase the robustness of the process at the level of the GRN dynamics? These and similar questions can be explored starting from the conservation data reported here and following a combination of theoretical and experimental approaches. A first approach to the role of duplications in the FOS-GRN can be found in Espinosa-Soto et al. (2004) for the case of the B-function genes in *Petunia*. This study showed that the FOS-GRN is dynamically robust to duplications.

In a study based on a comparative genomics approach, the quality of genome annotation is of major concern. The fact that putative annotation errors were detected recurrently in the same species gives support to the curational process followed, but it also suggests the need of more careful annotations in the genomes of *L. japonicus*, *O. sativa indica*, *P. trichocarpa*, *M. esculenta*, and *R. communis*. Future improvements in annotation quality may help the curational process in gene network conservation studies. Here, we report the conservation data for the FOS-GRN both before (supplementary table S5, Supplementary Material online) and after manual curation (fig. 2).

Selective Constraints in the FOS-GRN

It has been suggested that additional plant species, other than the experimental model species, should be included in molecular evolutionary studies to completely appreciate the conservation and evolvability of the regulatory network for flower development (Yang et al. 2011). Here, we show that the whole GRN controlling cell specification during early stages of flower development, when primordial floral organ cells are specified, has evolved under purifying selection. Unlike previous studies, we considered a wider range of angiosperms including both monocot and dicot species. Our results agree with previous conclusions: floral organ identity genes evolved under strong purifying selection. The evolution of the genes considered in the FOS-GRN is functionally constrained, as evidenced by the dN/dS ratios. We calculated an overall mean dN/dS of 0.124. From the 15 genes, 13 (86.66%) have a dN/dS value <0.15. Yang et al. recently reported the molecular evolutionary analysis of a group of 58 genes involved in flower development that includes all the genes that were analyzed in the present work, with the exception of EMF1 (Yang et al. 2011). Their analysis included only the species *A. thaliana* and *A. lyrata*. In their study, the authors report an average dN/dS value of 0.17 and interpret this result as evidence suggesting that these genes have overall evolved under purifying selection. On the other hand, the smaller average dN/dS value that we calculated for the 15 genes of the FOS-GRN (0.124) is based on a much wider range of species; and some of them are more distantly related than the two compared in the study of Yang et al. Furthermore, the calculated average value is highly influenced by the high dN/dS value corresponding to *EMF1* (0.39577). If we omit *EMF1* in the calculation, the average dN/dS is 0.1049864. This observation supports the conclusion

that the calculated dN/dS values are small and suggest that the FOS-GRN is functionally constrained. In order to find further support for our interpretation, we analyzed the dN and dS values previously reported for the whole-genome set of orthologous between *A. thaliana* and *A. lyrata* and calculated the average dN/dS value over the complete data set (see Materials and Methods). The calculated average dN/dS is 0.29; the complete empirical distribution is shown in supplementary figure S9*a*, Supplementary Material online. Using this whole-genome data set, we conducted a resampling experiment in order to calculate the likelihood of observing an average dN/dS value over a group of 15 genes equal or smaller to the one we report (0.124). The fraction of values from this distribution with a value equal or less than 0.124 was 0.00038. Hence, the encountered small value could be found in a random sample of the same size with a very small probability (P value = 0.00038). Supplementary figure S9*b*, Supplementary Material online, shows the distribution of simulated average dN/dS values. Considering that our dN/dS calculations are based on a set including more distant species, this empirical evidence strongly supports our claim that the reported average dN/dS of 0.124 is small.

When testing for evidence of positive selection as a force which could have fixed specific sites, using models that account for site class variability in dN/dS, we found that sites with a dN/dS >1 may exist only in *UFO*, *FT*, and *CLF* as evidenced by a marginal significant P value (before controlling for multiple tests) when comparing model M8 assuming positive selection with the null model M7 (see Materials and Methods). On the other hand, no single site in these proteins showed a high posterior probability when the Bayes' theorem was applied in order to identify potential targets of diversifying selection. Thus, in this study, both global and site varying models failed to detect any signature of positive selection for any codon of the FOS-GRN genes.

Unlike the above results, previous studies have found evidence of adaptive evolution acting at particular sites in some of the genes included in the FOS-GRN. Olsen et al. found evidence that suggests an adaptive mechanism behind the patterns of variation found on *TFL1* and *LFY*. These and similar studies (see, e.g., Olsen et al. 2002; Moore et al. 2005) are, in contrast to the present study, based on population genetic tests and data. Hence, these studies have captured the patterns of variation in these genes resulting from recent divergent evolution. Future studies should further investigate the microevolutionary process at play among the FOS-GRN genes. Some evidence at hand suggests that even for more recent divergences, floral organ identity genes will show evidence of strong purifying selection (Yang et al. 2011), but other flower transition genes seem to have been prone to positive selection as well (Martínez-Castilla and Alvarez-Buylla 2003); however, both selective forces are not mutually exclusive in any given gene.

Martínez-Castilla and Alvarez-Buylla (2003) focused on the Arabidopsis MADS-box gene family and found several sites within the MADS and K boxes, with high probabilities of having been fixed under positive selection, suggesting that these boxes may have played important roles in the acquisition of novel functions during recent events of MADS-box diversification. Here, through the analysis of alignments constructed on the basis of 1–1 orthologous relationships for distantly related angiosperm species, we did not find evidence of positive selection on such sites. Our result suggest that although adaptive evolution probably plays an important role during recent diversification events of the MADS-box gene family, a constrained evolution have prevailed upon the functionally established orthologous members across species which diverged more years ago. The question of whether or not the MADS-box gene family shows similar signs of adaptive evolution in species other than *A. thaliana* is open. This question, and its relevance for the phenotypic evolution of plants, is interesting given the complexity of the duplication events that have shaped the MADS-box gene family in angiosperms, as evidenced by the presence of multiple copies of flowering MADS-box genes found in several angiosperm species.

Selective Constraints and Functional Categories

Previous studies on floral genes in different populations of *A. thaliana* or different Arabidopsis species have also shown that floral organ identity genes evolved under strong purifying selection, but some flowering-time genes experienced relatively relaxed purifying selection and positive selection (Olsen et al. 2002; Moore et al. 2005). It has been suggested that selective constraints acting on genes of the same family are closely associated with their functions (Yang et al. 2011). The FOS-GRN includes genes which have been shown to be functionally associated with the promotion of flower meristem identity (*LFY*, *AP1*, *UFO*) or with floral organ identity (the ABC genes *AP1*, *AP2*, *AP3*, *PI*, *AG*). For historical and empirical reasons, the ABC genes have been qualified as having a prominent role in the process of cell fate and organ type specification during early flower development. Given this background information, the presence of a stronger functional constraint upon such genes in relation with the other interacting genes would be a reasonable hypothesis. Our results show that there is no significant difference between the molecular evolutionary parameters of these genes and the other genes in the FOS-GRN (supplementary table S7, Supplementary Material online), however. This suggests that the ABC genes have not been subject to a stronger functional constraint than the rest of the FOS-GRN genes, at least as evidenced by the differential rate of evolution analyses that we performed in this study. Instead, it seems that it is the whole regulatory module which is under a strong evolutionary constraint.

In contrast to the previous result, when molecular size and sequence conservation features were considered in addition to the dN/dS, it was possible to cluster the proteins into groups consistent with their functional roles. Specifically, an unsupervised model-based clustering analysis grouped the FOS-GRN proteins into three clusters consistent with their associated functions during inflorescence and flower development; and this consistency was assessed statistically

(see Results). Our results show that meristem and flower identity genes share similar molecular conservation features among them, whereas these are quite different from those observed in genes known to be involved in several other mechanisms with no apparent single prominent function. We interpret these results as evidence suggesting a constraint associated with the functional role of the genes. Although it is complicated to define rigorously a specific function for the individual components of complex molecular systems such as GRNs, given that no gene acts independently of their interacting partners or in a context-specific manner, our multivariate clustering approach uncovered a nontrivial pattern. Without any prior assumption about differences among the proteins, the methodology separated the genes in groups in a way consistent with the empirically known functions. Furthermore, the classification boundaries separating the clusters only merge in the case of the two groups in which some of their components are known to be associated with both functions (e.g., *AP1* is both a meristem and floral organ identity protein). Interestingly, it is only possible to uncover such a pattern when conservation measures other than evolutionary rates or sequence similarity were considered. The degree of conservation in sequence length seems to be relevant and closely associated with the molecular function. Finally, it is worth mentioning that the uncovered pattern is only obtained when considering several conservation features and not just a single evolutionary parameter or similarity measure.

Molecular Evolutionary Parameters and Network Architecture

Previous studies have suggested several approaches to test whether there is a relationship between network architecture and the molecular evolutionary parameters of the network's components (dN, dS, dN/dS): 1) the calculation of correlation coefficients between network topological measures of centrality and molecular evolutionary parameters (Montanucci et al. 2011), 2) the calculation of whether interacting nodes within a network have more similar values of the evolutionary parameters than noninteracting nodes (Alvarez-Ponce et al. 2009), and 3) the comparison of a matrix of pairwise shortest path distances between genes in the network and matrices of pairwise absolute differences in evolutionary parameters (Montanucci et al. 2011). Here, the three approaches were applied to the FOS-GRN, in addition to a regression-based modeling approach. Most of the above approaches assume that the architecture or topology of the network affects the molecular evolution of its nodes, and they implicitly assume then that such static network structure somehow is correlated to dynamical or functional modularity. Unlike previous network-level molecular evolutionary studies, we did not find a significant relationship between network architecture and the evolutionary parameters: 1) no significant correlation was found between the evolutionary parameters and the measures of centrality of the nodes, 2) analyses did not support the hypothesis that neighboring genes in the network share similar evolutionary constraints, and 3) regression coefficients did not support a relationship between the molecular evolutionary parameters and any of the nodes' topological features tested. This result suggests that the proteins of the FOS-GRN, although subject to purifying evolutionary forces, do not show any discernible pattern of association between the strength of constraint and the local structural properties within the network. This implies that the whole module is subject to similar molecular evolutionary constraints and/or the structural considerations do not have a functional or dynamical relevance that might have been important for the evolutionary constraints experienced by different nodes within the FOS-GRN. These results should be interpreted with caution, however, because of the small sample size. Statistical analysis has two goals that directly conflict. First is to find patterns in data. The second goal is a fight against apophenia, the human tendency to invent patterns in random data (Klemens 2008). In the context of GRNs, care should be taken when testing for the existence of relationships (or lack thereof) between node features and evolutionary patterns based on statistical analysis. The identification of "real patterns" could be limited by the size of the data set analyzed. Nonetheless, it is noteworthy that previous studies for small pathways/networks with a similar number of nodes as in the GRN analyzed here ($\leq$20 nodes) have found significant trends between topological and evolutionary parameters (see, e.g., Alvarez-Ponce et al. 2009; Fitzpatrick and O'Halloran 2012).

Given that we did find an association between conservation features of the genes—including evolutionary rates—and their functional role during flower development, and considering that the role of specific genes in the specification of meristem and floral identity has been probed during the analysis of the FOS-GRN as a dynamical system (Espinosa-Soto et al. 2004), we speculate that functional (dynamical), instead of topological, network properties, such as those associated with robustness, could be significantly associated with the molecular evolutionary constraints of the genes in the FOS-GRN reported here.

Overall, our results depict a general picture of the evolutionary pattern of the FOS-GRN where functional constraint better explains the evolution of its genes. The approach followed here provided new data relevant for the study of the evolution of the mechanisms at the molecular level that are behind organ identity during early flower development. Specifically, we have shown that 1) the FOS-GRN genes are conserved among 18 Angiosperm species; 2) a complex history of gene duplications seems to have been involved in the expansion of the network across angiosperms; 3) the whole FOS-GRN has evolved under purifying selection; 4) ABC floral organ identity genes do not show a significantly stronger evolutionary constraint than the other genes in the FOS-GRN; 5) an association between protein length and sequence conservation features, evolutionary rates, and functional category seems to prevail among the genes in the FOS-GRN; and 6) the FOS-GRN does not show any significant relationship between network architecture and the evolutionary parameters of its genes.

Materials and Methods

Sequence Data

The FOS-GRN described in Espinosa-Soto et al. (2004) and updated in Alvarez-Buylla et al. (2010) was used as study system; the corresponding genes are reported in supplementary table S1, Supplementary Material online. The identifiers of the genes involved in this network were obtained from the TAIR database (http://www.arabidopsis.org, last accessed November 24, 2013) and integrated into the workbench tool of the Plaza Comparative Genomics Platform (http://bioinformatics.psb.ugent.be/plaza/, last accessed November 24, 2013) (Proost et al. 2009).

After applying the PLAZA integrative method of orthologous genes finding (discussed later), both the sequence data of the genes of *A. thaliana* and the sequence data of the corresponding homologous genes were retrieved using the export functionality of the PLAZA'S workbench tool. This first data set corresponds to the FOS-GRN preliminary gene conservation set which includes those species with a sequenced and annotated genome and is represented as a conservation matrix and a list of corresponding gene identifiers in supplementary tables S2 and S3, Supplementary Material online, respectively. In order to reduce the probability of reporting the conservation of nonfunctional proteins, the preliminary data set was manually curated. For this purpose, erroneous automatic orthology designations were discarded, and those groups of adjacent gene annotations actually corresponding to different regions of a single gene were merged (discussed later). The final and corrected conservation data of the FOS-GRN proteins across angiosperms are reported in the form of a conservation matrix (fig. 1*a*) and its corresponding list of gene IDs (supplementary table S4, Supplementary Material online).

Homology Search

The PLAZA Comparative Genomics Platform offers an access point for plant comparative genomics centralizing genomic data produced by different genome sequencing initiatives (Proost et al. 2009). The PLAZA integrative method of orthologous genes integrates a complementary set of data types and methodologies in order to infer orthologous gene relationships based on the following sources of evidence: Orthologous gene families (ORTHO) inferred using OrthoMCL, Tree-based orthologs (TROG) inferred using tree reconciliation of the phylogenetic tree of a gene family, Best-Hits-and-Inparalogs (BHI) inferred from Blast hits against the PLAZA protein database, and Anchor points refer to gene-based colinearity between species. Using this tool, different homology relationship types can be considered: when a gene has no paralogs and only 1 ortholog (1–1), when a gene has 1 or more paralogs and only 1 ortholog (N–1), and the corresponding combinations for a total of four different orthology relationship: 1–1, N–1, 1–N, and M–N. In this work, the PLAZA integrative method was used to infer homology gene relationships for each protein in the FOS-GRN. The following settings were used: all orthologous relationship types were allowed, all evidence types were taken into account, and 18 plant species corresponding to the Phylum Angiospermae were included (see Results).

Manual Curation of Putative In-paralogs

As the degree and quality of annotation of whole-genome projects varies considerably among species, it is not adequate to rely only on automatic procedures, and instead, careful data set cleaning is necessary. Further manual curation to the reported gene groups after a homology analysis should be considered in order to reduce the likelihood of including nonfunctional proteins in other analyses. For each gene in the preliminary conservation data list (supplementary table S3, Supplementary Material online), the following information was extracted from PLAZA Comparative Genomics Platform: CDS sequence, protein sequence, chromosome, location (e.g., start, stop), length, and InterPro annotated protein domains. Given these data, some putative in-paralog genes were manually eliminated from the preliminary conservation data. On the other hand, the homology status of some genes was updated based on one or more of the following criteria: partial proteins (small size), lack of any of the protein domains of the orthologous gene in *A. thaliana*, neighboring genomic location, or low sequence alignment quality. The preliminary status of certain genes in the conservation data as multiple single paralogous copies in the same genome was modified to single copy orthologous genes, once it was realized that in many cases different boxes of the same open reading frame were sometimes annotated as different genes. Details of the manual curation process and sequence selection criteria are described in the supplementary text, Supplementary Material online.

Multiple Alignments and Phylogenetic Inference

All protein multiple sequence alignments (MSAs) were generated using the software CLUSTALW version 2.1 (Larkin et al. 2007). The software PAL2NAL (Suyama et al. 2006) was used to generate multiple codon alignments from the corresponding aligned protein sequences and the corresponding DNA coding sequences. For each orthologous group, a maximum likelihood phylogeny estimation was conducted using the software Phylm (Guindon and Gascuel 2003; Guindon et al. 2010) applying the nucleotide substitution model that best fits the data according to the Akaike information criterion. Details of the selected substitution models are provided on supplementary table S12, Supplementary Material online. Both phylogeny estimation and substitution model selection were conducted using the function phymltest of the package ape (Paradis et al. 2004) in the R statistical programming environment (www.R-project.org, last accessed November 24, 2013) as described in Paradis (2012).

Analysis of the Evolutionary Rates

The evolutionary parameters dN, dS, and dN/dS were estimated following a maximum likelihood procedure as implemented in the software codeml of the PAML package version

4.5 (Yang 2007). Due to the broad range of species considered for the conservation study, it was not possible to obtain reliable alignments for all 18 species for molecular evolutionary analysis. This analysis was then restricted to a representative group of species (*A. lyrata*, *A. thaliana*, *B. distachyon*, *G. max*, *M. esculenta*, *O. sativa*, *S. bicolor*, *T. cacao*, and *Z. mays*) to avoid bias in the dN/dS values—this decision was based on the manual inspection of the resulting alignments. All alignments are publicly available upon request. Only MSAs based on (putative) 1:1 ortholog sets were used. In the cases in which there were more than one gene copy in a given species, the gene with the most complete sequence or the one without any homogenization features (stop codons or frameshift mutations) was used. For each codon alignment, two tests of positive selection were performed. In order to test whether the assumption of positive selection fits better the data than the assumption of nearly neutral evolution, the model M2a was compared against the null model M1a through a likelihood ratio test (LRT). In a second test of positive selection, the models M7 (null model of neutral evolution), which assumes that dN/dS follows a (discrete) beta distribution among sites and M8 (positive selection model), which adds a class of dN/dS which can be greater than 1, were compared through an LRT. The false discovery rate and Bonferroni corrections for the multiple tests of positive selection were conducted using the function p.adjust of the stats package in the R statistical programming environment. In all the analyses, the F3×4 codon frequency model was used. Details of the LRT for each comparison are provided in supplementary table S5, Supplementary Material online. The strength of purifying selection was measured using the dN/dS values computed through the M0 model, which calculates rates encompassing all the branches of the tree and for the entire length of the sequence.

The dN and dS values reported for the whole-genome orthologous pairs between *A. thaliana* and *A. lyrata* were downloaded from the Ensemble Plant website (http://plants.ensembl.org/index.html, last accessed November 24, 2013) using the BioMart platform for data retrieval. The corresponding dN/dS values and their statistics were calculated over the complete data set, omitting missing data (a total of 22,531 values). The empirical distribution is shown in supplementary figure S9*a*, Supplementary Material online. A resampling experiment was conducted using the complete set of dN/dS values as follows: a large number of gene groups of size 15 (100,000) were randomly generated, the dN/dS average value was calculated for each group, and the distribution obtained values was used to estimate the likelihood of observing an average value equal or smaller than the one calculated for the FOS-GRN (0.124). The simulated distribution is shown in supplementary figure S9*b*, Supplementary Material online.

Gene Conservation Features

The pairwise distances from protein MSAs were calculated using the function dist.ml of the package phangorn (Schliep 2011) with the default parameters. In the case of DNA codon MSAs, a matrix of pairwise distances was computed using the dist.dna function of the package ape (Paradis et al. 2004) with the default parameters. To obtain a final scalar conservation feature, the corresponding means were calculated and used as a summary statistics. The coefficient of variation in protein size of each protein over all species was calculated as a measure of the degree of conservation (variation) in molecular size. All the calculations discussed in this section were conducted using the R statistical programming environment.

Genes Clustering and Function

Hypothesis tests of statistical difference of the evolutionary parameters between the ABC floral organ identity genes and the other genes in the FOS-GRN were conducted following a nonparametric method (Kruskal-Wallis test). A model-based clustering analysis was conducted using the molecular and sequence conservation features in table 2 (last four columns) as an input feature matrix. Intuitively, the goal of clustering is to assign points that are similar to the same cluster and to ensure that points that are dissimilar are in different clusters. The analysis was conducted as implemented in the function Mclust of the mclust package version 4.1 (Fraley et al. 2012). This procedure fits a Gaussian finite mixture model to the data through an EM algorithm. The best model is selected according to the Bayesian information criterion. The clustering procedure was evaluated using the functional categories of the genes as an external form of data for validation. The clustering was then compared with the labels using as summary statistic the Rand index, which measures the fraction of clustering decisions that are correct (Murphy 2012). The Rand index was calculated using the function cluster_similarity of the package clusteval (http://cran.r-project.org/web/packages/clusteval/, last accessed November 24, 2013). In order to assess the statistical significance of the clustering, the frequentist sampling distribution of a standard summary statistic that quantifies the fraction of clustering decision that are correct was computed using a bootstrap method. The Rand index was used as a summary statistic (Murphy 2012). Specifically, a character vector corresponding to the clustering output was permuted a large number of times (n = 1,000,000) and compared each time with the labels vector using the Rand index. The obtained sampling distribution was used to calculate the probability of observing a Rand index value equal or greater than the one observed when comparing the original output of the clustering analysis with the labels vector. Both model-based clustering analysis and clustering evaluation were conducted in the R statistical programming environment.

Evolutionary Rates and Network Architecture

The measures of centrality describe numerically the topological importance of a node in a graph, given its structure. For each gene (node) in the FOS-GRN, the following measures of centrality were calculated: degree (number of nodes it is connected to), closeness (reciprocal of the average distance to all other nodes), betweenness (fraction of all shortest paths that pass through it), and eccentricity (maximum distance

from it to all other nodes). All network topological computations were conducted using the igraph package (Csardi and Nepusz 2006). Two analyses were conducted in order to test for the association of the evolutionary parameters of the genes and their topological features within the network. 1) Spearman correlation coefficients were calculated between each evolutionary parameter given by the model M0 (dN, dS, and dN/dS) and each topological features. 2) Simple linear regression models were fitted using each evolutionary parameter as response variable and each topological feature as predictor.

It was also investigated whether genes that are interacting in the FOS-GRN have related values of the evolutionary parameters. For this purpose, two additional analyses were conducted. In the first analysis, following Alvarez-Ponce et al. (2009), the average absolute difference of the value of the evolutionary parameters between interacting components in the network was calculated and then used as an statistic in a simulation (sampling) procedure in order to assess how frequently it is expected to observe this or a smaller value in an ensemble of similar but random networks. Specifically, 100,000 networks each with the same number of nodes and interactions were generated, and the statistic was calculated for each of these networks. The estimated distribution of the statistic over the ensemble of networks was then used to calculate the probability of observing a value equal or smaller than that calculated in the FOS-GRN. A Gaussian density function with parameters estimated from the data (mean = 0.0713 and standard deviation = 0.0128) was also fitted from the observed simulated data and used for probability calculations. In the second analysis, following Montanucci et al. (2011), a matrix of pairwise shortest path distances between the nodes (path distance matrix) and three matrices of absolute pairwise gene differences in each of the evolutionary parameters were computed. Each of these last matrices was then compared with the path distance matrix through standardized Mantel tests using the ecodist package. All the analyses discussed in this section were conducted using the R statistical programming environment.

Supplementary Material

Supplementary figures S1–S9 and tables S1–S12 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

E.R.A.B. acknowledges the support of the Miller Institute for Basic Research in Science, University of California, Berkeley, while spending a sabbatical leave in the lab of Chelsea Specht. The authors acknowledge technical support of Rigoberto V. Pérez-Ruiz and logistical and administrative help of Diana Romo. This article constitutes a partial fulfillment of the graduate program Doctorado en Ciencias Biomédicas of the Universidad Nacional Autónoma de México, UNAM in which J.D.-V. developed this project. This work was supported by grants from CONACYT, Mexico: 180098 (to E.R.A.B.) from PAPIIT-UNAM, IN203113-3 (to E.R.A.B.).

References

Agrafioti I, Swire J, Abbott J, Huntley D, Butcher S, Stumpf MP. 2005. Comparative analysis of the *Saccharomyces cerevisiae* and *Caenorhabditis elegans* protein interaction networks. *BMC Evol Biol.* 5:23.

Alvarez-Buylla ER, Azpeitia E, Barrio R, Benitez M, Padilla-Longoria P. 2010. From ABC genes to regulatory networks, epigenetic landscapes and flower morphogenesis: making biological sense of theoretical approaches. *Semin Cell Dev Biol.* 21:108–117.

Alvarez-Buylla ER, Chaos A, Aldana M, et al. (11 co-authors). 2008. Floral morphogenesis: stochastic explorations of a gene network epigenetic landscape. *PLoS One* 3:e3626.

Alvarez-Buylla ER, Pelaz S, Liljegren SJ, Gold SE, Burgeff C, Ditta GS, De Pouplana LR, Martínez-Castilla L, Yanofsky MF. 2000. An ancestral MADS-box gene duplication occurred before the divergence of plants and animals. *Proc Natl Acad Sci U S A.* 97:5328–5333.

Alvarez-Ponce D. 2012. The relationship between the hierarchical position of proteins in the human signal transduction network and their rate of evolution. *BMC Evol Biol.* 12:192.

Alvarez-Ponce D, Aguadé M, Rozas J. 2009. Network-level molecular evolutionary analysis of the insulin/TOR signal transduction pathway across 12 *Drosophila* genomes. *Genome Res.* 19:234–242.

Alvarez-Ponce D, Aguadé M, Rozas J. 2011. Comparative genomics of the vertebrate insulin/TOR signal transduction pathway: a network-level analysis of selective pressures. *Genome Biol Evol.* 3: 87–101.

Alvarez-Ponce D, Fares MA. 2012. Evolutionary rate and duplicability in the *Arabidopsis thaliana* protein-protein interaction network. *Genome Biol Evol.* 4:1263–1274.

Becker A, Theissen G. 2003. The major clades of MADS-box genes and their role in the development and evolution of flowering plants. *Mol Phylogenet Evol.* 29:464–489.

Bowers JE, Chapman BA, Rong J, Paterson AH. 2003. Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. *Nature* 422:433–438.

Bremer B, Bremer K, Chase M, Fay M, Reveal J, Soltis D, Soltis P, Stevens P. 2009. An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG III. *Bot J Linn Soc.* 161:105–121.

Casals F, Sikora M, Laayouni H, Montanucci L, Muntasell A, Lazarus R, Calafell F, Awadalla P, Netea MG, Bertranpetit J. 2011. Genetic adaptation of the antibacterial human innate immunity network. *BMC Evol Biol.* 11:202.

Coen ES, Meyerowitz EM. 1991. The war of the whorls: genetic interactions controlling flower development. *Nature* 353:31–37.

Cork JM, Purugganan MD. 2004. The evolution of molecular genetic pathways and networks. *Bioessays* 26:479–484.

Csardi G, Nepusz T. 2006. The igraph software package for complex network research. *InterJournal, Complex Systems* 1695:5.

Espinosa-Soto C, Padilla-Longoria P, Alvarez-Buylla ER. 2004. A gene regulatory network model for cell-fate determination during *Arabidopsis thaliana* flower development that is robust and recovers experimental gene expression profiles. *Plant Cell* 16: 2923–2939.

Fraley C, Raftery AE, Murphy TB, Scrucca L. 2012. MCLUST version 4 for R: normal mixture modeling for model-based clustering, classification, and density estimation. Technical Report no. 597, Department of Statistics, University of Washington.

Fraser HB, Hirsh AE, Steinmetz LM, Scharfe C, Feldman MW. 2002. Evolutionary rate in the protein interaction network. *Science* 296(5568):750–752.

Fitzpatrick DA, O'Halloran DM. 2012. Investigating the relationship between topology and evolution in a dynamic nematode odor genetic network. *Int J Evol Biol.* 2012(2012):548081.

Guindon S, Dufayard J, Lefort V, Anisimova M, Hordijk W, Gascuel O. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol.* 59:307–321.

Guindon S, Gascuel O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol.* 52: 696–704.

Ha M, Kim ED, Chen ZJ. 2009. Duplicate genes increase expression diversity in closely related species and allopolyploids. *Proc Natl Acad Sci U S A.* 106:2295–2300.

Hahn MW, Conant GC, Wagner A. 2004. Molecular evolution in large genetic networks: does connectivity equal constraint? *J Mol Evol.* 58: 203–211.

Hahn MW, Kern AD. 2005. Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks. *Mol Biol Evol.* 22(4):803–806.

Huang S, Kauffman S. 2009. Complex gene regulatory networks-from structure to biological observables: cell fate determination. In: Meyers RA, editor. Encyclopedia of complexity and systems science. Berlin: Springer. p. 1180–1293.

Invergo BM, Montanucci L, Laayouni H, Bertranpetit J. 2013. A system-level, molecular evolutionary analysis of mammalian phototransduction. *BMC Evol Biol.* 13:52.

Jovelin R, Phillips PC. 2009. Evolutionary rates and centrality in the yeast gene regulatory network. *Genome Biol.* 10:R35.

Klemens B. 2008. Modeling with data: tools and techniques for scientific computing. Princeton (NJ): Princeton University Press.

Krizek BA, Fletcher JC. 2005. Molecular mechanisms of flower development: an armchair guide. *Nat Rev Genet.* 6:688–698.

Larkin MA, Blackshields G, Brown NP, et al. (13 co-authors). 2007. Clustal W and Clustal X version 2.0. *Bioinformatics* 23:2947–2948.

Lavagnino N, Serra F, Arbiza L, Dopazo H, Hasson E. 2012. Evolutionary genomics of genes involved in olfactory behavior in the *Drosophila melanogaster* species group. *Evol Bioinform Online.* 8:89–104.

Lawton-Rauh AL, Alvarez-Buylla ER, Purugganan MD. 2000. Molecular evolution of flower development. *Trends Ecol Evol.* 15:144–149.

Lemos B, Bettencourt BR, Meiklejohn CD, Hartl DL. 2005. Evolution of proteins and gene expression levels are coupled in *Drosophila* and are independently associated with mRNA abundance, protein length, and number of protein-protein interactions. *Mol Biol Evol.* 22(5):1345–1354.

Martínez-Castilla L, Alvarez-Buylla ER. 2003. Adaptative evolution in the *Arabidopsis* MADS-box gene family inferred from its complete resolved phylogeny. *Proc Natl Acad Sci U S A.* 100(23):13407–13412.

Mendoza L, Alvarez-Buylla ER. 1998. Dynamics of the genetic regulatory network for *Arabidopsis thaliana* flower morphogenesis. *J Theor Biol.* 193(2):307–319.

Montanucci L, Laayouni H, Dall'Olio GM, Bertranpetit J. 2011. Molecular evolution and network-level analysis of the N-glycosylation metabolic pathway across primates. *Mol Biol Evol.* 28:813–823.

Moore RC, Grant SR, Purugganan MD. 2005. Molecular population genetics of redundant floral-regulatory genes in *Arabidopsis thaliana*. *Mol Biol Evol.* 22:91–103.

Murphy K. 2012. Machine learning: a probabilistic approach. Cambridge (MA): MIT Press.

Olsen KM, Womack A, Garrett AR, Suddith JI, Purugganan MD. 2002. Contrasting evolutionary forces in the *Arabidopsis thaliana* floral developmental pathway. *Genetics* 160:1641–1650.

Paradis E. 2012. Analysis of phylogenetics and evolution with R. New York: Springer.

Paradis E, Claude J, Strimmer K. 2004. APE: analyses of phylogenetics and evolution in R language. *Bioinformatics* 20:289–290.

Preston JC, Kellogg EA. 2007. Conservation and divergence of APETALA1/FRUITFULL-like gene function in grasses: evidence from gene expression analyses. *Plant J.* 52:69–81.

Proost S, Van Bel M, Sterck L, Billiau K, Van Parys T, Van de Peer Y, Vandepoele K. 2009. PLAZA: a comparative genomics resource to study gene and genome evolution in plants. *Plant Cell* 21:3718–3731.

Purugganan MD. 1997. The MADS-box floral homeotic gene lineages predate the origin of seed plants: phylogenetic and molecular clock estimates. *J Mol Evol.* 45:392–396.

Purugganan MD, Rounsley SD, Schmidt RJ, Yanofsky MF. 1995. Molecular evolution of flower development: diversification of the plant MADS-box regulatory gene family. *Genetics* 140:345–356.

Purugganan MD, Suddith JI. 1999. Molecular population genetics of floral homeotic loci: departures from the equilibrium-neutral model at the APETALA3 and PISTILLATA genes of *Arabidopsis thaliana*. *Genetics* 151:839–848.

Riechmann JL, Meyerowitz EM. 1997. MADS domain proteins in plant development. *J Biol Chem.* 378:1079.

Sanchez-Corrales YE, Alvarez-Buylla ER, Mendoza L. 2010. The *Arabidopsis thaliana* flower organ specification gene regulatory network determines a robust differentiation process. *J Theor Biol.* 264: 971–983.

Schliep KP. 2011. phangorn: phylogenetic analysis in R. *Bioinformatics* 27: 592–593.

Suyama M, Torrents D, Bork P. 2006. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res.* 34:W609–W612.

Villarreal C, Padilla-Longoria P, Alvarez-Buylla ER. 2012. General theory of genotype to phenotype mapping: derivation of epigenetic landscapes from N-node complex gene regulatory networks. *Phys Rev Lett.* 109:118102.

Wellmer F, Riechmann JL. 2010. Gene networks controlling the initiation of flower development. *Trends Genet.* 26:519–527.

Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol.* 24:1586–1591.

Yang L, Chun-Ce G, Gui-Xia H, Hong-Yan S, Hong-Zhi K. 2011. Evolutionary pattern of the regulatory network for flower development: insights gained from a comparison of two *Arabidopsis* species. *J Syst Evol.* 49:528–538.

Yang Y, Zhang F, Ge S. 2009. Evolutionary rate patterns of the Gibberellin pathway genes. *BMC Evol Biol.* 9:206.

Yang Z, Bielawski JP. 2000. Statistical methods for detecting molecular adaptation. *Trends Ecol Evol.* 15:496–503.

CellPress
PARTNER JOURNAL

Molecular Plant
Research Article

XAANTAL2 (*AGL14*) Is an Important Component of the Complex Gene Regulatory Network that Underlies *Arabidopsis* Shoot Apical Meristem Transitions

Rigoberto V. Pérez-Ruiz[1,4], Berenice García-Ponce[1,4,*], Nayelli Marsch-Martínez[1,5], Yamel Ugartechea-Chirino[1], Mitzi Villajuana-Bonequi[1,6], Stefan de Folter[2], Eugenio Azpeitia[1,7], José Dávila-Velderrain[1], David Cruz-Sánchez[1], Adriana Garay-Arroyo[1], María de la Paz Sánchez[1], Juan M. Estévez-Palmas[1] and Elena R. Álvarez-Buylla[1,3,*]

[1]Instituto de Ecología, Universidad Nacional Autónoma de México, 3er Circuito Exterior s/no, Junto al Jardín Botánico, and Centro de Ciencias de la Complejidad Ciudad Universitaria, Coyoacán 04510, México D.F., Mexico

[2]Laboratorio Nacional de Genómica para la Biodiversidad (Langebio), Centro de Investigación y de Estudios Avanzados del Instituto Politécnico Nacional (CINVESTAV-IPN), Km. 9.6 Carretera Irapuato - León, AP 629, 36821 Irapuato, Guanajuato, Mexico

[3]University of California, 431 Koshland Hall, Berkeley, CA 94720, USA

[4]These authors contributed equally to this article.

[5]Present address: Departamento de Biotecnología y Biquímica, Centro de Investigación y de Estudios Avanzados del Instituto Politécnico Nacional (CINVESTAV-IPN), Km. 9.6 Libramiento Norte, Carr. Irapuato-León, 36821 Irapuato, Guanajuato, Mexico

[6]Present address: Max Planck Institute for Plant Breeding Research, D-50829 Cologne, Germany

[7]Present address: INRIA Project-Team Virtual Plants/CIRAD/INRA, UMR AGAP Campus St Priest - BAT 5, CC 05018, 860 rue de St Priest, 34095 Montpellier Cedex 5, France

***Correspondence: Berenice García-Ponce (bgarcia@ecologia.unam.mx), Elena R. Álvarez-Buylla (eabuylla@gmail.com)**

http://dx.doi.org/10.1016/j.molp.2015.01.017

ABSTRACT

In *Arabidopsis thaliana*, multiple genes involved in shoot apical meristem (SAM) transitions have been characterized, but the mechanisms required for the dynamic attainment of vegetative, inflorescence, and floral meristem (VM, IM, FM) cell fates during SAM transitions are not well understood. Here we show that a MADS-box gene, *XAANTAL2* (*XAL2*/*AGL14*), is necessary and sufficient to induce flowering, and its regulation is important in FM maintenance and determinacy. *xal2* mutants are late flowering, particularly under short-day (SD) condition, while *XAL2* overexpressing plants are early flowering, but their flowers have vegetative traits. Interestingly, inflorescences of the latter plants have higher expression levels of *LFY*, *AP1*, and *TFL1* than wild-type plants. In addition we found that XAL2 is able to bind the *TFL1* regulatory regions. On the other hand, the basipetal carpels of the *35S::XAL2* lines lose determinacy and maintain high levels of *WUS* expression under SD condition. To provide a mechanistic explanation for the complex roles of *XAL2* in SAM transitions and the apparently paradoxical phenotypes of *XAL2* and other MADS-box (*SOC1*, *AGL24*) overexpressors, we conducted dynamic gene regulatory network (GRN) and epigenetic landscape modeling. We uncovered a GRN module that underlies VM, IM, and FM gene configurations and transition patterns in wild-type plants as well as loss and gain of function lines characterized here and previously. Our approach thus provides a novel mechanistic framework for understanding the complex basis of SAM development.

Key words: *XAL2*/*AGL14*, MADS-box, *TFL1*, SAM transitions, floral reversion, gene regulatory networks, epigenetic landscape modeling

Pérez-Ruiz R.V., García-Ponce B., Marsch-Martínez N., Ugartechea-Chirino Y., Villajuana-Bonequi M., de Folter S., Azpeitia E., Dávila-Velderrain J., Cruz-Sánchez D., Garay-Arroyo A., Sánchez M.P., Estévez-Palmas J.M., and Álvarez-Buylla E.R. (2015). *XAANTAL2* (*AGL14*) Is an Important Component of the Complex Gene Regulatory Network that Underlies *Arabidopsis* Shoot Apical Meristem Transitions. Mol. Plant. **8**, 796–813.

Published by the Molecular Plant Shanghai Editorial Office in association with Cell Press, an imprint of Elsevier Inc., on behalf of CSPB and IPPE, SIBS, CAS.

INTRODUCTION

Unraveling the molecular genetic mechanisms that underlie cell transitions and plasticity is a fundamental issue in developmental biology. Different cell states (e.g., proliferative, differentiated, transdifferentiated, or reprogrammed) are correlated to different combinations of gene activation (Sugimoto et al., 2011). Such gene configurations, and the transitions among them, emerge from complex regulatory networks (Álvarez-Buylla et al., 2010a, 2010b). Plants enable *in vivo* analyses of the molecular genetic mechanisms underlying such cell plasticity and dynamics of stem cells that remain active during their complete life cycle within meristems.

At the shoot apical meristem (SAM) the transition from a vegetative to a reproductive state is crucial, with direct fitness implications (Roux et al., 2006). Molecular genetic approaches have uncovered a complex gene regulatory network (GRN) underlying *Arabidopsis* SAM development (Srikanth and Schmid, 2011; Andrés and Coupland, 2012). Genetic screenings for mutant plants with altered bolting time under contrasting environmental conditions (Koornneef et al., 1991) have uncovered the components of flowering transition pathways in response to: photoperiod (Putterill et al., 1995; Suárez-López et al., 2001; An et al., 2004), gibberellins (gibberellic acid [GA]; Blázquez et al., 1998; Blázquez and Weigel, 2000; Porri et al., 2012), non-optimal growth temperature over 4°C (Blázquez et al., 2003; Halliday et al., 2003; Balasubramanian et al., 2006; Lee et al., 2007), vernalization (Michaels and Amasino, 1999; Sheldon et al., 2000; Michaels et al., 2003), or internal developmental cues (Koornneef et al., 1991; Simpson, 2004; Wu and Poethig, 2006).

Many of the genes that participate in floral transition are MADS-box genes (Gramzow et al., 2010). Some of them, such as *SUPPRESSOR OF OVEREXPRESSION OF CONSTANS 1* (*SOC1*), respond to more than one condition, and these have been called integrators (Blázquez and Weigel, 2000; Lee et al., 2000; Moon et al., 2003; Wang et al., 2009; Lee and Lee, 2010). Detailed functional characterization revealed that flowering transition pathways converge in the regulation of *LEAFY* (*LFY*) and *APETALA1* (*AP1*), via SOC1–AGAMOUS-LIKE 24 (AGL24) heterodimer, SQUAMOSA PROMOTER BINDING PROTEIN-LIKE (SPL3) or FLOWERING LOCUS T–FLOWERING LOCUS D (FT–FD) complex, at the founding cells of the floral meristem (FM), thus establishing a new identity distinct from the inflorescence meristem (IM). The FM later sub-differentiates into the floral organs (Schultz and Haughn, 1991; Weigel et al., 1992; Abe et al., 2005; Yamaguchi et al., 2009).

Gene expression configurations that characterize the IM and FM identities, in addition to the floral organ primordia, have started to be recovered and explained with dynamic GRN mechanistic models, as attractors or steady states (Espinosa-Soto et al., 2004; Álvarez-Buylla et al., 2010a; van Mourik et al., 2010; Kaufmann et al., 2011; Jaeger et al., 2013). Such mechanistic explanations are still lacking for normal and altered cell-fate transitions at the SAM in wild-type plants, and for certain MADS-box overexpression lines (Yu et al., 2004; Ferrario et al., 2004; Liu et al., 2007; Fornara et al., 2008).

The coexistence and, at the same time, the clear distinction of IM and FM suggest a common underlying dynamic multi-stable mechanism. Some genes have been identified as critical markers of each of these SAM cellular identities, while others are shared among them. Distinction between IM and FM depends on the mutual repression of floral meristem identity genes, such as *LFY*, *AP1*, and *CAULIFLOWER* (*CAL*), and IM genes, particularly *TERMINAL FLOWER1* (*TFL1*), an important regulator of inflorescence development (Shannon and Meeks-Wagner, 1991; Alvarez et al., 1992; Weigel et al., 1992; Bowman et al., 1993; Shannon and Meeks-Wagner, 1993; Gustafson-Brown et al., 1994; Chen et al., 1997; Ohshima et al., 1997; Ratcliffe et al., 1998, 1999; Ferrándiz et al., 2000; Parcy et al., 2002). *TFL1* encodes a phosphatidylethanolamine-binding protein (PEBP) that is transcribed in the center of the IM, but the protein moves to other cells where *AP1* and *LFY* are down-regulated (Bradley et al., 1997; Conti and Bradley, 2007). *tfl1* is an early flowering mutant with a determinate inflorescence due to the ectopic expression of *LFY* and *AP1* in the IM (Shannon and Meeks-Wagner, 1991; Schultz and Haughn, 1993; Gustafson-Brown et al., 1994; Mandel and Yanofsky, 1995; Liljegren et al., 1999). Conversely, single and double mutants of *LFY* and *AP1* acquire inflorescence-like structures because of the ectopic expression of *TFL1* (Huala and Sussex, 1992; Bowman et al., 1993; Bradley et al., 1997; Ratcliffe et al., 1998, 1999; Benlloch et al., 2007).

Recent data show that the tight spatial and temporal regulation of the components of the GRN underlying the transition to flowering is also involved in FM identity and maintenance (Liu et al., 2009; Posé et al., 2012). In this sense, genes such as *SOC1*, *AGL24*, and *SHORT VEGETATIVE PHASE* (*SVP*), known to participate in the regulation of flowering transition by regulating *LFY* in the case of the first two genes (Lee et al., 2008; Liu et al., 2008), and SVP in collaboration with FLOWERING LOCUS C (FLC) by repressing *SOC1* and *FT* (Hartmann et al., 2000; Lee et al., 2007; Li et al., 2008), are also important during the first two stages of flower development (Gregis et al., 2009; Liu et al., 2009). At these stages, *SOC1*, *AGL24*, and *SVP* help to prevent the premature expression of the B and C genes (Gregis et al., 2006, 2009; Liu et al., 2009). Moreover SOC1, AGL24, SVP, and SEP4 with AP1 repress the expression of *TFL1* in the FM (Liu et al., 2013). At stage 3 of FM development, *AGL24* and *SVP* are repressed by LFY and AP1, leading to further differentiation and determinacy (Yu et al., 2004; Liu et al., 2007). Meanwhile, expression of *SOC1* and *FRUITFULL* (*FUL*, another MADS-box gene) in the IM is important to repress secondary vascular growth (Melzer et al., 2008). Therefore, *SOC1*, *AGL24*, *SVP*, and *FUL* are important in both flowering transition, and floral and inflorescence meristems identity and maintenance.

Additional evidence for the common underlying multi-stable and non-linear GRN for SAM states and transitions is the fact that several of the aforementioned MADS-domain proteins are involved in multiple SAM states and transitions (Smaczniak et al., 2012), sometimes with apparently paradoxical functions. The overexpression of some MADS-box genes, such as *AGL24* or *SOC1* and their homologs, induce early flowering by up-regulating *LFY* and *AP1* (Lee et al., 2000; Yu et al., 2002; Michaels et al., 2003; Lee et al., 2008), but at the same time produce flowers with vegetative characteristics that resemble the *ap1* mutant with elongated carpels, especially

under short-day (SD) condition (Irish and Sussex, 1990; Bowman et al., 1993; Borner et al., 2000; Ferrario et al., 2004; Masiero et al., 2004; Yu et al., 2004; Liu et al., 2007; Trevaskis et al., 2007; Fornara et al., 2008). The phenomenon known as "floral reversion" has been also described in heterozygous *lfy*, *ap1*, *ap2*, and *agamous* (*ag*) mutants, suggesting that these genes repress this process and favor FM determinacy (Battey and Lyndon, 1990; Okamuro et al., 1993, 1996, 1997). There is no explanation or mechanistic model to account for the permanence of inflorescence characteristics when *LFY* and *AP1* are prematurely expressed in the MADS-box overexpression lines.

XAANTAL2 (*XAL2*/*AGL14*) is a MADS-box gene preferentially expressed in the root (Rounsley et al., 1995; Garay-Arroyo et al., 2013). The name *XAANTAL2* was given because *xal2* mutants have short roots similar to those of *xaantal1*/*agl12* (Tapia-López et al., 2008; Garay-Arroyo et al., 2013). Here, we report that *XAL2* is also a key player in SAM cell identities and transitions. It promotes flowering and presents similar loss and gain of function phenotypes such as *AGL24* and *SOC1*. We also show that overexpression of *XAL2*, *SOC1*, and *AGL24* are able to up-regulate *TFL1*, thus explaining, at least in part, the prevalence of vegetative traits, even if *AP1* and *LFY* are prematurely expressed, supporting that *XAL2* is also important for FM maintenance. Here, we propose a dynamic GRN and epigenetic landscape (EL) models (Álvarez-Buylla et al., 2008, 2010b; Villarreal et al., 2012) that integrate our data with previous results to provide a mechanistic and dynamic framework to understanding normal and altered cell fates and transitions at the *Arabidopsis* SAM. This model thus provides a mechanistic explanation for apparently paradoxical data for other loss and gain of function phenotypes (Borner et al., 2000; Ferrario et al., 2004; Masiero et al., 2004; Yu et al., 2004; Liu et al., 2007; Trevaskis et al., 2007; Fornara et al., 2008) allowing the integration of additional components.

RESULTS

XAL2 Promotes Flowering Transition

XAL2 is a member of the TM3/SOC1 clade, belonging to the type II MADS-box genes (Álvarez-Buylla et al., 2000; Martínez-Castilla and Álvarez-Buylla, 2003; Parenicová et al., 2003; Smaczniak et al., 2012). Except for *XAL2* (Garay-Arroyo et al., 2013), all other members of this clade have been identified as activators of flowering transition (Lee et al., 2000; Moon et al., 2003; Schmid et al., 2003; Schönrock et al., 2006; Dorca-Fornell et al., 2011). Given the role of all other members of *SOC1* clade, we hypothesized that *XAL2* could also be involved in flowering and tested two *xal2* alleles under four conditions: long-day (LD) and SD photoperiods, vernalization plus LD, and GA_3 treatment plus SD. In addition, we generated double mutants using the *xal2-2* allele (which has less somatic *En*-excision rates than *xal2-1*) and *soc1-6*, *agl24-4*, and *ful-7* mutants, because SOC1, AGL24, and FUL proteins interact with XAL2 in the yeast two-hybrid system, suggesting that they form dimers (de Folter et al., 2005; van Dijk et al., 2010).

Under LD condition both *xal2* alleles (Garay-Arroyo et al., 2013) showed a subtle but significant delay in bolting time (Figure 1A and 1B and Supplemental Table 1). Under the same condition, *soc1-6* was epistatic over *xal2-2*, while *xal2-2* and *ful-7* had a slightly additive effect on bolting time compared with the parental plants. No differences were observed in the *xal2-2 agl24-4* double mutant with respect to single mutants (Figure 1B). Interestingly, the rosette leaf number (RLN) did not always coincide with the bolting time phenotype (Figure 1B and 1C). In fact, *xal2-1* and *xal2-2* alleles and *xal2-2 ful-7* have the same number of leaves as wild-type plants under LD condition, while *xal2-2 soc1-6* double mutants had fewer leaves than *soc1-6* (Figure 1B and Supplemental Table 1).

Under SD condition both *xal2* alleles are remarkably delayed compared with wild-type plants and only *xal2-2 soc1-6* plants showed an additive bolting time phenotype in comparison with both parentals (Figure 1C and Supplemental Table 1). However, *xal2-2* was epistatic over *agl24-4* and *ful-7* mutants under this condition (Figure 1C). Unexpectedly, the *xal2-2 soc1-6* RLN is lower than in both parental lines (Figure 1C). Therefore, it seems that *XAL2* effects on bolting time and rosette leaf development are partially independent. We also found that cauline leaf number is diminished in *xal2-2* only under SD condition and is epistatic over *soc1-6*, *agl24-4*, and *ful-7* (Supplemental Figure 1A).

Since GA plays a relevant role in flowering under SD, we tested the effect of this hormone in all mutants. GA application partially suppressed flowering phenotypes under SD condition in all cases except for *xal2-2 soc1-6* (Figure 1C and 1D). Interestingly, 62% of the *xal2-2 soc1-6* plants grown under SD condition were unable to flower after 117 days after sowing (DAS), and none of them flowered after GA treatment (88 DAS), thus suggesting that *XAL2* and *SOC1* additively participate in GA response during flowering transition. To explore how the impairment of GA response in *xal2-2 soc1-6* affects GA homeostasis, we assayed two GA biosynthesis genes (*GA20OX1* and *2*) and a catabolic one (*GA2OX1*; Rieu et al., 2008) at 14 DAS, when most of the flowering time genes are up-regulated under LD condition. Our results in the double mutant showed up-regulation of *GA20OX1* compared with *xal2-2* and down-regulation of *GA2OX1* compared with wild-type plants (Supplemental Figure 2A). This finding suggests a compensatory mechanism in which the plant tries to make up for reduced GA responses by producing more GA. Further analysis should be performed to clarify the role of *XAL2* in relation to *SOC1* in GA homeostasis during flowering transition.

Overall, our results for single and double mutants indicate that both *xal2* alleles have a delayed bolting time compared with wild-type plants under all conditions tested, except for vernalization treatment (Figure 1A–1D and Supplemental Table 1). To further explore the role of *XAL2* in flowering transition and to uncover possible redundancies of this gene with other related MADS-box genes, we generated several *35S::XAL2* lines and selected three of them that showed the highest levels of *XAL2* transcript accumulation (Supplemental Figure 2B) and similar phenotypes among them (see description in the following paragraphs). In Figure 1E and Supplemental Figure 1B we show that *35S::XAL2* line (9T4) has a similar early bolting time and fewer rosette and cauline leaves in comparison with wild-type plants, under both LD and SD condition. Therefore, *XAL2* is

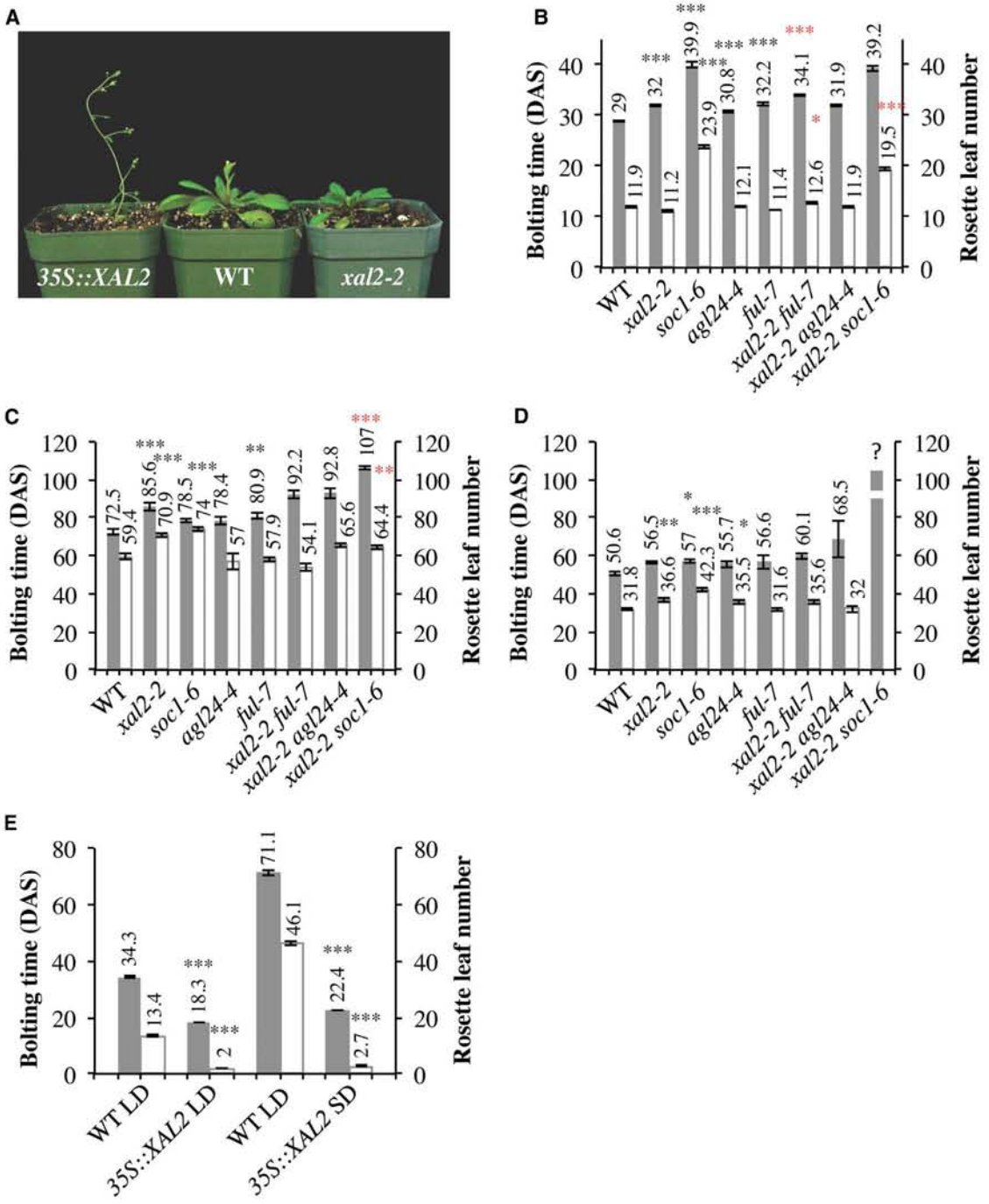


Figure 1. *XAL2* Participates in Flowering Transition.

(A) The mutant allele *xal2-2* and the overexpression line *35S::XAL2* are late and early flowering compared with wild-type (WT) plants, respectively.

(B) Flowering time of double mutant plants *xal2-2 ful-7*, *xal2-2 agl24-4*, and *xal2-2 soc1-6* compared with parental and WT plants grown under long-day (LD) condition, showing that *soc1-6* is epistatic over *xal2-2*. DAS, days after sowing.

(C) The same plants grown under short-day (SD) condition showed that the *xal2-2 soc1-6* double mutant plants have an additive effect compared with the parental and WT plants.

(D) GA_3 application mostly suppressed the late flowering phenotype of all genotypes. Note that none of the *xal2-2 soc1-6* double mutant plants flowered after 88 DAS.

(E) Overexpression of *XAL2* is sufficient to induce a similar early bolting time phenotype under LD and SD conditions.

Flowering transition was analyzed as the bolting time (gray bars) expressed in DAS and the rosette leaf number (white bars) as mean ± standard error (n = 35–42 plants under LD and n = 16–23 under SD and SD + GA). Lines with statistically significant differences compared with WT plants (black asterisks) or single mutants (red asterisks) are indicated as $^{*}P < 0.05$, $^{**}P < 0.01$, and $^{***}P < 0.001$ according to one-way analysis of variance (ANOVA) following Tukey's multiple comparison test.

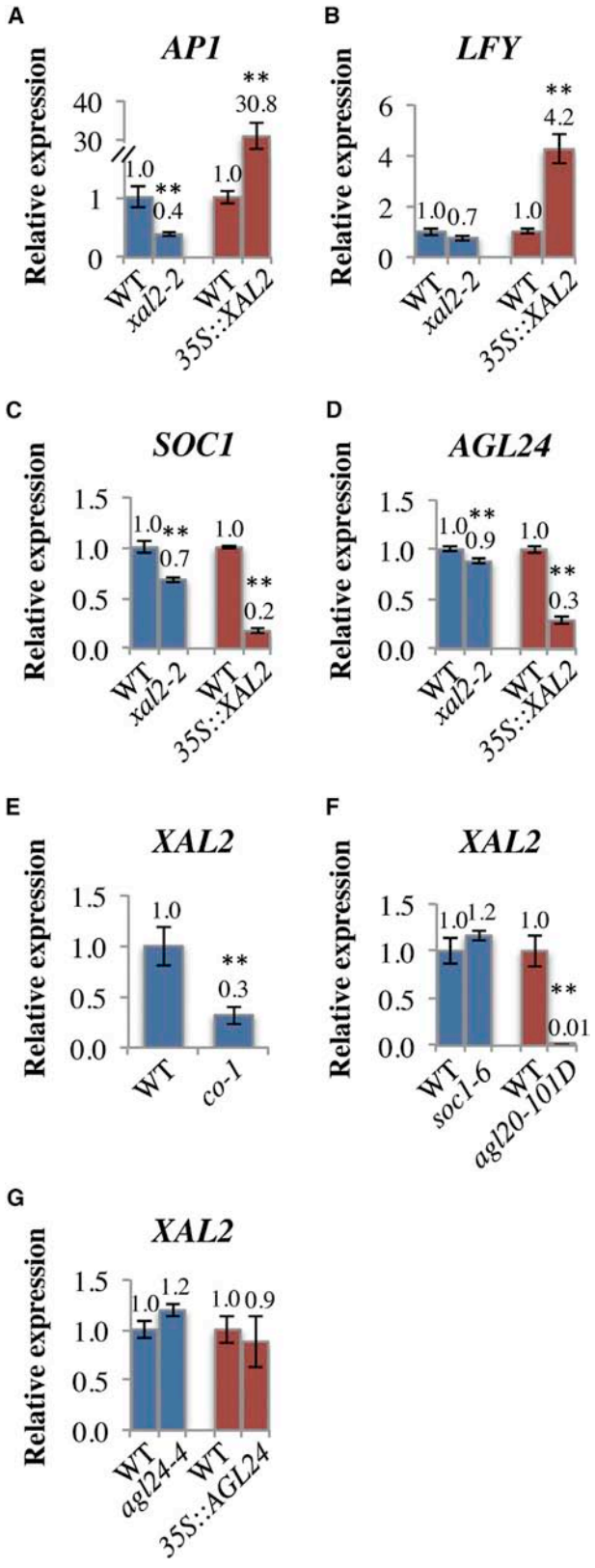


Figure 2. *XAL2* Regulation in the Flowering Gene Regulatory Network (GRN) under LD condition.

(A) XAL2 positively regulates *AP1*.

(B) *LFY* is up-regulated by XAL2 only in the overexpression line.

(C and D) *SOC1* **(C)** and *AGL24* **(D)** are down-regulated in *xal2-2*, and are repressed in the *XAL2* overexpression line.

(E) CO positively regulates *XAL2*.

(F) Overexpression of *SOC1* represses the expression of *XAL2*, but no significant difference in the latter was observed in *soc1-6* with respect to WT.

(G) AGL24 does not regulate *XAL2*.

Relative mRNA accumulation from three biological replicates were obtained from 14 DAS seedlings (blue bars) and 10 DAS plants (red bars) grown under LD condition. Data are shown as mean ± standard error. Statistical significance (**$P < 0.01$) was evaluated using the Mann–Whitney test.

sufficient to induce early flowering independently of photoperiod, and may participate in the IM to FM transition. These results confirm that *XAL2* is a key component of the GRN that controls flowering transition.

XAL2 Is Part of the GRN that Induces *AP1* during Floral Transition

We then analyzed the role of *XAL2* in the flowering transition GRN using quantitative RT–PCR. In agreement with the flowering transition phenotypes observed in Figure 1, we found that *AP1* expression was down-regulated in *xal2-2* (Figure 2A) while *LFY* did not show significant repression (Figure 2B). In contrast, both genes were up-regulated in the *XAL2* overexpression line (Figure 2A and 2B). *SOC1* and *AGL24* were also down-regulated in *xal2-2*, indicating that XAL2 positively regulates both genes. Surprisingly, *XAL2* overexpression drastically repressed *SOC1* and *AGL24* (Figure 2C and 2D). Therefore, it is possible that the early flowering phenotype observed in the *XAL2* overexpression line is due to an up-regulation of *LFY* and *AP1* and that this is partially independent of SOC1-AGL24.

XAL2 is down-regulated in *constans-1* mutant (*co-1*; Han et al., 2008), indicating that CO positively regulates *XAL2* when plants are grown under LD condition (Figure 2E). *XAL2* transcript accumulation in *soc1-6* (Wang et al., 2009) and *agl24-4* was unaffected (Figure 2F and 2G). Thus, at the transcriptional level, *XAL2* is regulated by CO and positively regulates *SOC1* and *AGL24*. Interestingly, when we analyzed *XAL2* accumulation in the *SOC1* (*agl20-101D*; Lee et al., 2000) and *AGL24* (Yu et al., 2002) overexpression lines, we found that *XAL2* was strongly repressed only in *agl20-101D* (Figure 2F and 2G). This suggests that *XAL2* and *SOC1* overexpression lines induce early flowering independently of one another.

In summary, the RT–PCR results indicate that *XAL2* is an important component of the GRN that regulates *AP1*, is under the control of CO, and participates in the up-regulation of *SOC1* and *AGL24* upon floral transition.

XAL2 Participates in Flower Meristem Maintenance and Determinacy

To address the role of *XAL2* in FM development, we analyzed its spatio-temporal expression pattern with *in situ* hybridization at different FM stages. *XAL2* expression appears very early at the flanks of the IM in the anlagens upon the transition to flowering (Figure 3A). Subsequently, *XAL2* expression levels increase in the first and second stages of the FM (Smyth et al., 1990), and

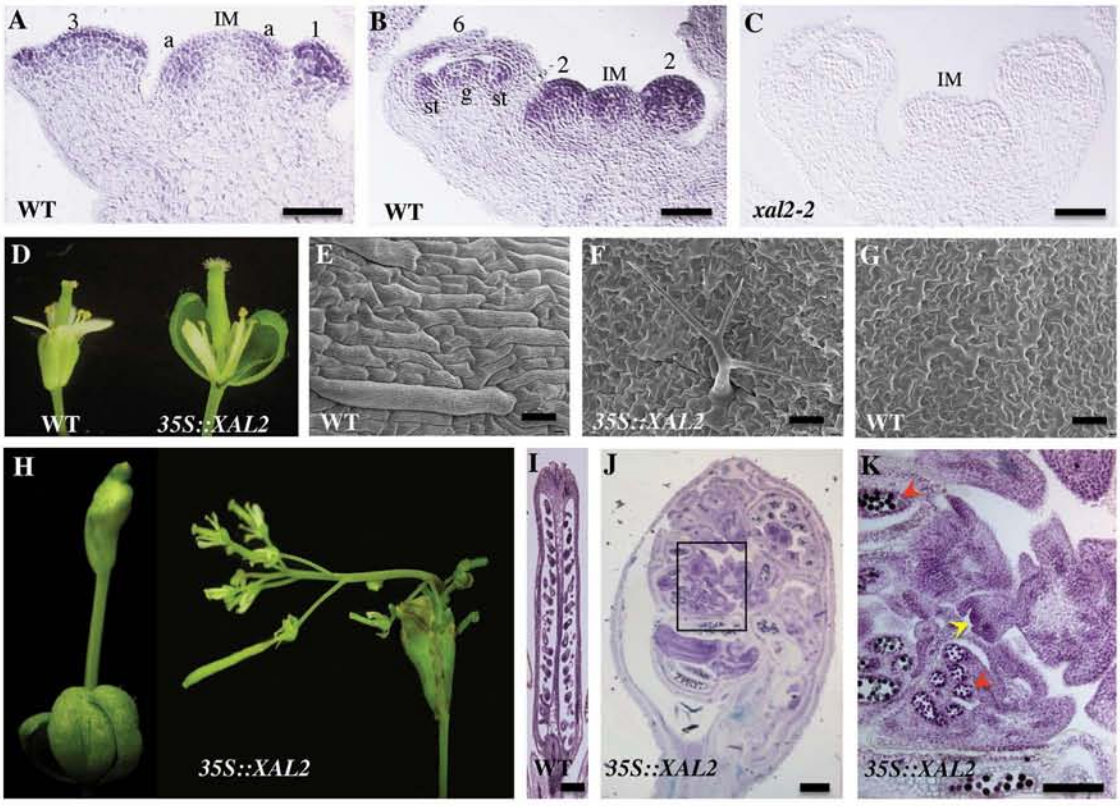


Figure 3. *XAL2* Spatial and Temporal Expression in WT SAM during the Transition to Flowering and *XAL2* Overexpression Floral Phenotypes.

(A–C) mRNA *XAL2 in situ* hybridization in WT and *xal2-2* inflorescences. **(A)** *XAL2* is detected in the anlagen (a), stage 1 of the floral meristem (FM), and L1 and L2 layers of FM stage 3. **(B)** *XAL2* accumulates at stage 2 of the FM and at the inflorescence meristem (IM). Later in FM development (stage 6), *XAL2* is also detected in the stamen (st) and gynoecium (g) primordia. **(C)** As a negative control, no signal was detected when *XAL2* antisense probe was used in the *xal2-2* mutant.

(D–G) Floral phenotype of the *35S::XAL2* compared with WT grown under LD condition. **(D)** *35S::XAL2* sepals are larger than WT sepals, and scanning electron micrographs show that the cellular identity of the *35S::XAL2* sepals **(F)** is more similar to WT leaf cells **(G)** than to WT sepal cells **(E)**, including the presence of trichomes **(F)**.

(H) Under SD condition, early arising carpels of the *35S::XAL2* plants elongate and inflorescences develop inside them.

(I–K) Longitudinal toluidine blue-stained sections confirmed that flowers at different developmental stages can be observed inside the *35S::XAL2* carpels compared with a similar stage of WT carpels based on ovule development **(I)**. **(K)** Magnification of the rectangle in **(J)** shows a FM at stage 4 of development (yellow arrowhead) and pollen grains (red arrowheads) inside the *XAL2* overexpression carpel.

Scale bars correspond to 50 μm **(A–C)**, 20 μm **(E–G)**, and 500 μm **(I–K)**.

at stage 3, *XAL2* is restricted to the L1 and L2 layers (Figure 3A and 3B). Later on, *XAL2* is expressed in the gynoecium and stamen primordia at stage 6 (Figure 3B). Interestingly, *XAL2* mRNA is also detected in the IM periphery (Figure 3B). We used *xal2-2* mutant as a negative control to rule out cross-hybridization of our probe with the closely related *AGL19* mRNA (Figure 3C).

The *XAL2* spatio-temporal expression pattern is similar to that of *AGL24* and *SVP* (Yu et al., 2004; Liu et al., 2007; Gregis et al., 2009). It has been reported that *SOC1*, in addition to the latter two genes, are important during the first two stages of FM development, but are repressed at stage 3 for proper subsequent FM differentiation (Yu et al., 2004; Gregis et al., 2006; Liu et al., 2007; Gregis et al., 2009; Liu et al., 2009). Furthermore, LFY and AP1 (particularly the latter) repress *AGL24* and *SVP* at FM stage 3 (Yu et al., 2004; Liu et al., 2007). Coincidentally, we found that *XAL2* accumulation is higher in the *ap1-15* (Ng and Yanofsky, 2001) and *ap1-1 cal-5* mutants (Ferrándiz et al., 2000) and is down-regulated in the *tfl1-2* mutant, in which the IM is converted into FM (Supplemental Figure 3A; Shannon and Meeks-Wagner, 1991; Alvarez et al., 1992). In agreement, an opposite pattern of expression for *LFY* was detected in these mutants (Supplemental Figure 3A). Therefore, AP1 and CAL probably repress *XAL2* in the FM at stage 3, as occurs with *AGL24* and *SVP*.

As already explained, *XAL2* overexpression induces early flowering with the production of very few rosette leaves (Figure 1A and 1E). It is noteworthy that cauline leaves in these lines are rounder and larger, similar to rosette leaves (Figure 5A), and flowers have leaf-like traits, such as large sepals that remain indehiscent after fertilization (Figures 3D and 5D), and sepal cells with a morphology reminiscent of wild-type leaf cells (Figure 3E–3G). These phenotypes are similar to those reported for *SOC1*, *AGL24*, and their homolog overexpression lines (Borner et al.,

2000; Michaels et al., 2003; Ferrario et al., 2004; Masiero et al., 2004; Yu et al., 2004; Liu et al., 2007; Trevaskis et al., 2007; Fornara et al., 2008). Interestingly, in the *35S::XAL2* under SD condition, early arising basipetal carpels deformed (stages 14–17 of flower development; Smyth et al., 1990; Roeder and Yanofsky, 2006) and a whole inflorescence grew from inside, bearing new fertile flowers (Figure 3H, 3J, and 3K). The flowers that develop from these indeterminate carpels attain different developmental stages, from FM (stage 4 in the picture) up to flowers with mature ovules and pollen grains (Figure 3I and 3K). These results indicate that correct spatio-temporal control of *XAL2* expression is fundamental for normal FM cell differentiation and determinacy.

In conclusion, *XAL2* overexpression accelerates the transition to flowering by terminating the vegetative phase prematurely, but at the same time the flowers produced in these lines show leaf-like traits. In addition, under non-inductive flowering conditions, overexpression of *XAL2* prevents FM determinacy, leading to what appears to be cell reprogramming with some carpel cells functioning as IM cells.

XAL2 Overexpression Positively Regulates *TFL1* and *WUS* Expression, and Directly Binds to the *TFL1* Regulatory Sequences

To unravel the molecular basis of the *XAL2* overexpression phenotypes in which this gene up-regulates *AP1* and *LFY* (Figure 2A and 2B), and at the same time yields flowers with some *ap1* mutant characteristics (Figure 3D and 3F; Irish and Sussex, 1990; Bowman et al., 1993), we hypothesized that an inflorescence identity gene capable of repressing *AP1* could be involved. Hence, we analyzed the expression of *TFL1* in *xal2-2* and the 35S::*XAL2* line, and found that it was down- and up-regulated in these lines, respectively (Figure 4A). Furthermore, to establish whether XAL2 is able to directly bind to *TFL1* regulatory sequences, we performed a chromatin immunoprecipitation (ChIP) experiment using a *35S::GFP-XAL2* line. In Figure 4B we show that three different *TFL1* regulatory regions containing CArG boxes are enriched (III, V, and VI) within the 5′ promoter and the intergenic region downstream of the 3′ stop codon of the *TFL1* gene. These results strongly support that under constitutive expression, *XAL2* directly binds to *TFL1* regulatory sequences. Since *SOC1* and *AGL24* overexpression phenotypes are similar to those of *35S:XAL2*, we further analyzed *TFL1* transcript accumulation in these two lines. *TFL1* was also up-regulated in *agl20-101D* and *35S::AGL24* (Figure 4C).

We have already shown that *AP1* is up-regulated in the *XAL2* overexpression line at 10 DAS (Figure 2A), but we wanted to be sure that the *ap1*-like phenotype was not due to its down-regulation at different developmental stages. We performed an *AP1* expression time course from 8 to 14 DAS plants, and at all time points analyzed *35S:XAL2* plants showed higher levels of *AP1* than wild-type plants (Supplemental Figure 3B). Hence, we can conclude that leaf-like traits of the *35S::XAL2* flower organs are not due to decreased levels of *AP1*. Also, *SOC1* and *AGL24* overexpressors showed up-regulation of *AP1* as expected (Figure 4D). Thus, overexpression of *XAL2*, *SOC1*, or *AGL24* is able to up-regulate both *TFL1* and *AP1* and cause early flowering, and at the same time yield flowers with leaf-like cell types.

FM cells stop proliferating after the carpels are formed in wild-type flowers (Mizukami and Ma, 1997). *WUSCHEL* (*WUS*), a key gene involved in the identity of stem cells in the SAM, is repressed in the central zone of the FM at stage 6 (Lenhard et al., 2001; Lohmann et al., 2001; Sun et al., 2009). Since we observed that in non-inductive flowering conditions FM determinacy is lost and a new inflorescence emerges from inside of early arising carpels in the *XAL2* overexpression lines (Figure 3H, 3J, and 3K), we hypothesized that *WUS* is persistently expressed in these lines. Therefore, we analyzed *WUS* and *TFL1* mRNA accumulation in these carpels compared with wild-type carpels at similar developmental stages. Indeed, we found that both genes were up-regulated in the *XAL2* overexpression line (Figure 4E and 4F), confirming that when *XAL2* is de-regulated, some of the molecular components that are important for IM and FM identity and determinacy are also altered.

To test whether the floral "reversion" phenotype of the *35S::XAL2* line was due to higher competence of *TFL1* over *AP1*, we crossed it to a *35S::AP1* plant to test whether the excess of *AP1* could counteract *TFL1* (Figure 5; Mandel and Yanofsky, 1995). As expected, both lines partially complemented each other's phenotypes in the double overexpressor line grown under SD condition, resulting in plants with smaller cauline leaves and flowers with reduced sepals compared with the *35S::XAL2* parental plant (Figure 5A, 5C, 5D, and 5F). On the other hand, the conversion of inflorescences into solitary flowers, typical of the *35S::AP1* line, disappeared (Figure 5B and 5C). Although both parental lines were early flowering, the bolting time of the *35S::AP1 35S::XAL2* line was the same as for *35S::AP1* plants, but the double overexpressor line had an intermediate number of rosette leaves with respect to both parental lines (Figure 5G). Interestingly, the double overexpressor had fewer swollen carpels compared with the *XAL2* overexpression line (Figure 5H), indicating that the indeterminacy observed in the *35S::XAL2* FM (Figure 3H, 3J, and 3K) was almost recovered when *AP1* was increased.

GRN and EL Modeling for *XAL2* Interactions under LD Condition: A Mechanistic Dynamic Explanation for *XAL2*, *SOC1*, and *AGL24* Overexpression Phenotypes

Our data uncover a complex set of interactions and roles for *XAL2* in SAM transitions. To provide an integrative, system-level, dynamic and mechanistic explanation for our results, a GRN modeling approach is required. We integrated the evidence of this work, together with previously reported information (Supplemental Table 2), to uncover a necessary and sufficient set of components and interactions (i.e., dynamic GRN module) that recover observed patterns of gene expression in vegetative meristem (VM), IM, and FM cells in wild-type plants. These patterns correspond to the expected set of steady-state gene configurations to which such wild-type GRN should converge (Figure 6 and Table 1), and can be validated if it also explains what we observe in the analyzed loss and gain of function lines.

We formalized experimental data as logical functions following previous studies (see the Methods section and Supplemental

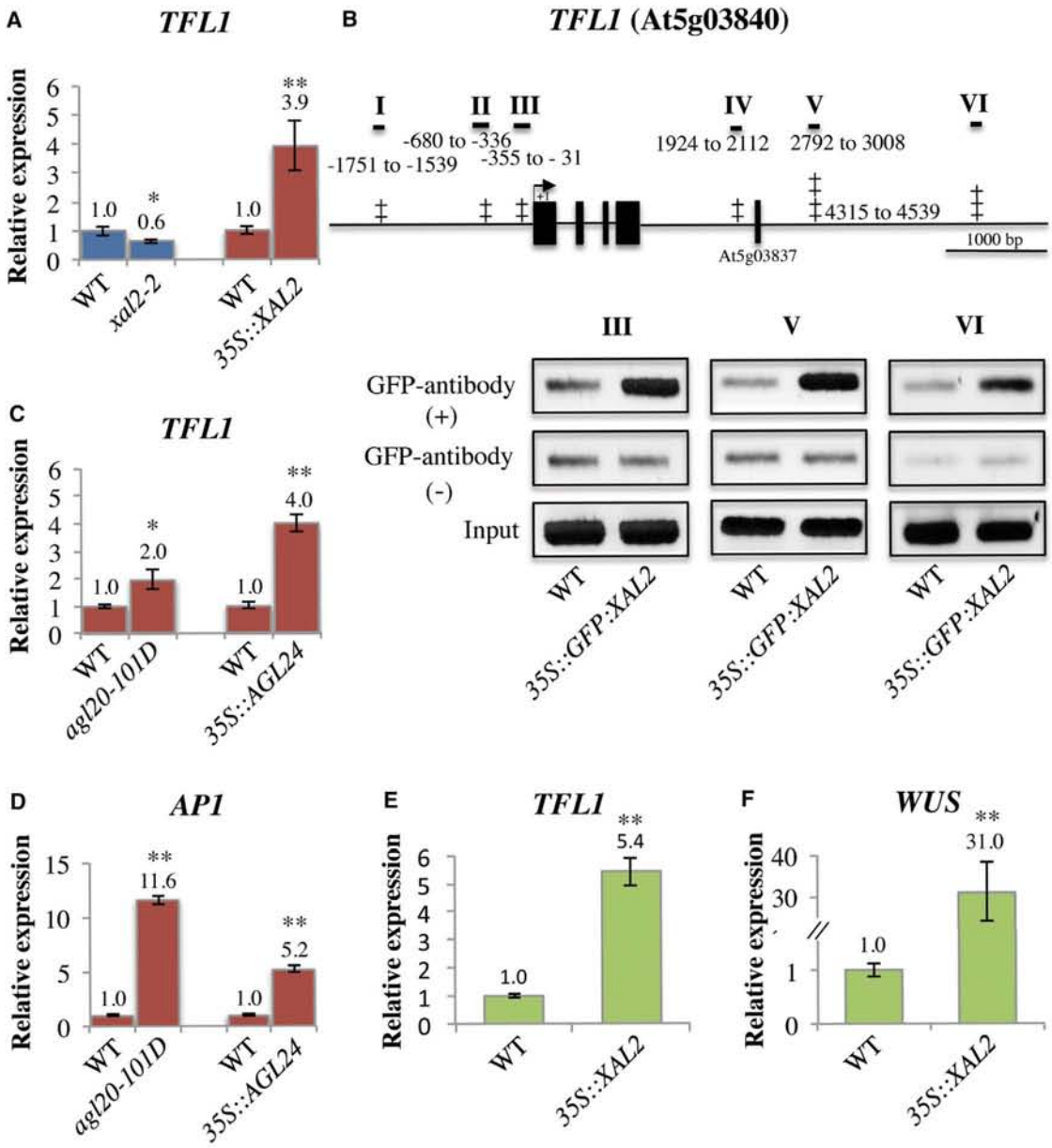


Figure 4. *XAL2* is a Positive Regulator of *TFL1*.

(A) *TFL1* relative RNA accumulation is down-regulated in *xal2-2* and up-regulated in the *XAL2* overexpression line.

(B) Chromatin immunoprecipitation (ChIP) assay was performed to examine *in vivo* binding of XAL2 to the *TFL1* regulatory regions in the *XAL2* overexpression line. Top panel shows a schematic diagram of the *TFL1* locus, indicating in roman numerals the regions amplified by PCR after DNA immunoprecipitation. Primers flanking the CArG boxes (+) and their positions relative to the *TFL1* transcriptional start site are indicated. The bottom panel shows DNA fragments corresponding to *TFL1* III, V, and VI regions enriched in the *35S::GFP:XAL2* plants after ChIP with a GFP antibody.

(C and D) Up-regulation of *TFL1* **(C)** and *AP1* **(D)** in the *SOC1* (*agl20-101D*) and *35S::AGL24* overexpression lines.

(E and F) Higher RNA accumulation levels of both *TFL1* **(E)** and *WUS* **(F)** were detected in first arising carpels of the *35S::XAL2* plants compared with WT plants grown under SD condition.

Quantitative RT–PCR was performed with RNA extracted from 14 DAS seedlings (blue bars), 10 DAS seedlings (red bars), and carpels with similar ovule stage development (green bars). Data in **(A)** and **(C–F)** are shown as mean ± standard error. Statistical significance (*$P < 0.05$, **$P < 0.01$) was evaluated using Mann–Whitney test.

Table 3). We were able to recover a wild-type GRN model under LD condition that integrates the data presented in this work and the necessary and sufficient set of additional components and interactions from the literature, to recover the expected steady states for VM, IM, and FM for the genes considered (Figure 6C and Table 1). Loss of function lines were simulated by turning the corresponding gene to "0" during the complete simulation, while the overexpression lines were simulated by turning the corresponding gene to "2." The proposed GRN is validated because, as expected by the observed phenotypes, all single loss of function mutants qualitatively recovered the same set of steady states as wild-type GRNs, while gain of function GRN

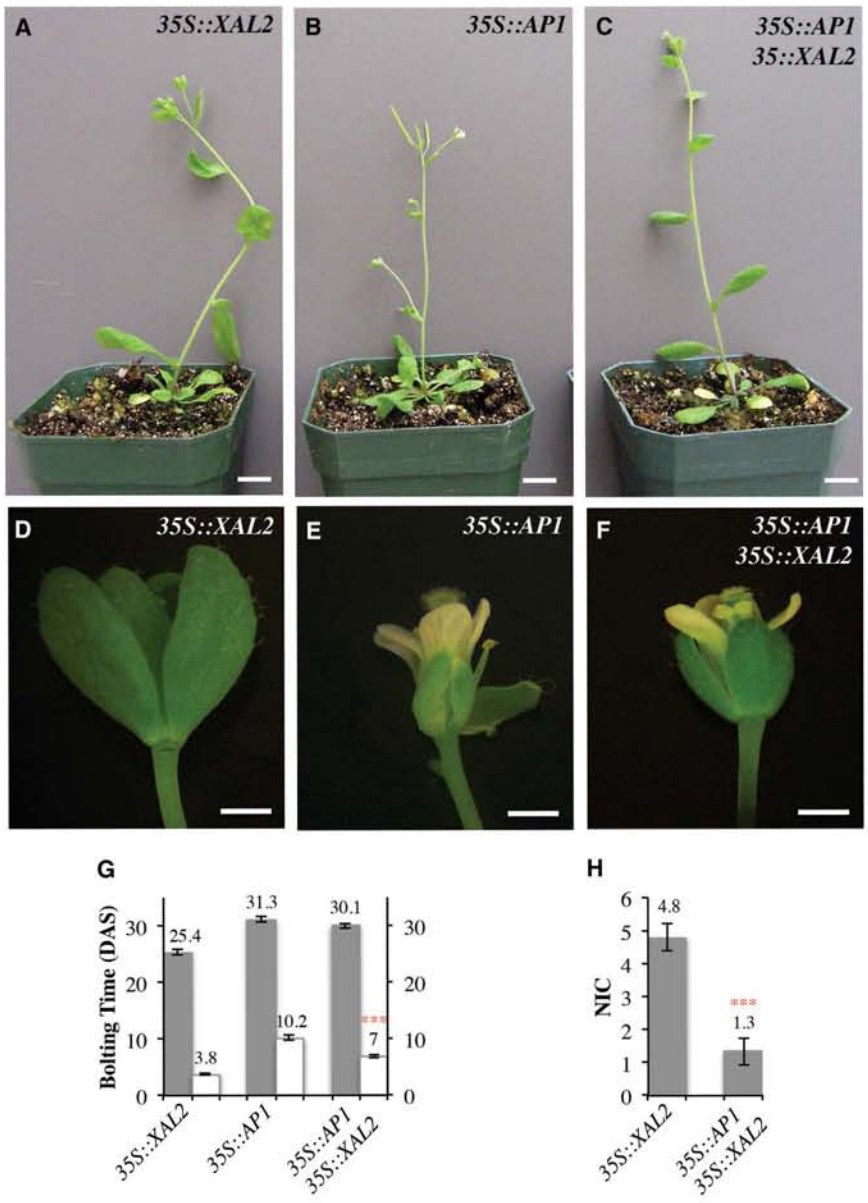


Figure 5. Complementation Analysis of the *35S::XAL2* and the *35S::AP1* Phenotypes in the Double Overexpressor Plants Grown under SD Condition.

(A and D) *XAL2* overexpression plants have large cauline leaves similar to rosette leaves **(A)** and flowers with large sepals that persist after fertilization **(D)**.

(B and E) *35S::AP1* plants show a determinate growth in which each pedicel gives rise from two to three terminal flowers **(B)**. Flowers of the *35S::AP1* are similar to those of WT **(E)**.

(C and F) Determinate growth of the *35S::AP1* line is complemented in the double overexpressor *35S::AP1 35S::XAL2* plants. On the other hand, the cauline leaves phenotype of the *35S::XAL2* is complemented to WT in this line **(A–C)**. Sepals of the double overexpressor line are partially complemented, resulting in sepals that are much smaller than the *35S::XAL2* sepals **(D and F)**. Scale bars correspond to 1 cm **(A–C)** and 2 mm **(D–F)**.

(G and H) The double overexpressor plants **(G)** have the same bolting time as the *35S::AP1* line, but have an intermediate rosette leaf number compared with parental plants. The number of indeterminate carpels (NIC) along the shoot axis of the double overexpressor **(H)** is also reduced compared with the *35S::XAL2* line. Bars correspond to standard error from average (n = 26–32 plants). Statistical significance with respect to parental plants (***P < 0.001; red asterisks) was evaluated according to one-way ANOVA following Tukey's multiple comparison test **(G)** or Mann–Whitney test **(H)**.

simulations recovered wild-type steady states, plus a new steady state with an IM/FM mixed cell identity (Figure 6C and Table 1). Indeed, experimental data has shown that *soc1*, *agl24*, and *xal2* single mutants do not modify cell identities but only flowering time, which cannot be simulated with this version of the model. On the other hand, *XAL2*, *SOC1*, or *AGL24* overexpression not only modifies flowering time, but also produces some flowers with inflorescence-like characteristics. Coincidentally, our model suggests that such flowers have some cells with a mixed IM/FM identity.

To gain further insight into how the alteration in the expression of *SOC1*, *AGL24*, and especially *XAL2* modify SAM cell transitions, we propose an EL analysis similar to that reported by Álvarez-Buylla et al. (2008) (Figure 7 and Supplemental Figure 4). Such analysis addresses whether the set of components and interactions considered in the uncovered GRN module in Figure 6B also underlies the observed temporal pattern of transition among cell types in wild-type and other lines (steady states): VM > IM > FM. Importantly, this type of model can discriminate between two hypotheses: the observed leaf-like structures in flowers of the overexpressors is due to a reversion from FM cells to IM cells, or in these lines a new type of steady state with mixed identity (IM/FM) appears during SAM development. Thus, this and the GRN modeling provide a mechanistic explanation for the apparently paradoxical phenotype of the overexpressors.

We thus performed a stochastic simulation of the proposed GRN model to propose a model for a population of cells at the SAM (see Supplemental Methods). Since VM cells are the first to attain their fate in wild-type, all cells were assumed to be in this state at initial conditions. Thus, in the vector with the proportion of cells in each GRN steady state for the dynamic stochastic equation, VM was set to 1 and the rest to 0 (Figure 7A–7D). This equation was iterated to follow the changes in the probability of reaching each one of the other steady states over time. The graph clearly shows how the trajectory for each of the steady states' probability reaches its maximum at a given time. In accordance with biological observations, the results show that the most probable sequence of cell attainment is VM > IM > FM

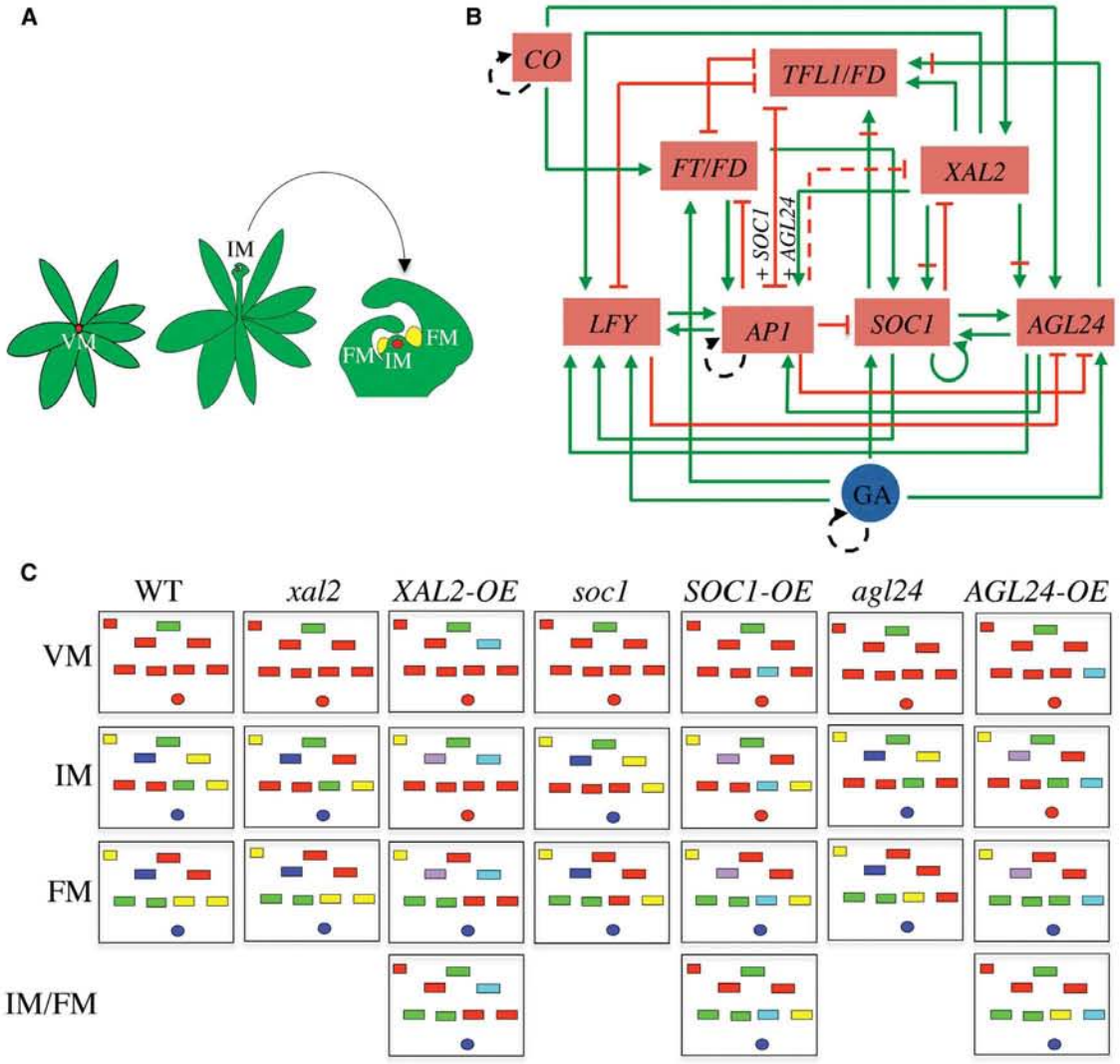


Figure 6. Model for the *XAL2* Regulatory Network Module during SAM Development and Its Steady States for the WT, Loss and Gain of Function *XAL2*, *SOC1*, and *AGL24* Lines.

(A) Schematic representation of SAM transitions from a vegetative (VM) to inflorescence (IM) and floral (FM) meristem states.

(B) GRN showing the interactions uncovered in this paper and published results (see Supplemental Table 2). Arrows (green) and bar-lines (red) indicate induction and repression, respectively. In some cases, we discovered that the sign of the interaction inferred changed depending if the loss or gain of function lines were being tested (regulation of XAL2, SOC1, and AGL24 over some of their targets). Dotted lines represent predictions of regulations that need further verification. In the case of GA and *CO*, the positive feedback loops are introduced because their upstream regulators that keep them turned on were not considered in the model proposed here. AP1 plus SOC1 or AGL24 indicate protein dimers that repress *TFL1* (Liu et al., 2013).

(C) A schematic representation of the network in **(B)** is used to represent the steady states achieved by this model under the different lines considered (columns). In each row, the steady states corresponding to the VM, IM, FM, or the novel IM/FM state recovered in the overexpressors. The components of the network are shown by squares or a circle (GA) that are turned on/off in each of the steady-state configurations being considered. The colors correspond to the activation state of the node in each case: red = 0; green = 1; yellow = 0 or 1; purple = 1 or 2; light blue = 2; and dark blue = 0, 1, or 2.

in wild-type plants (Figure 7A). In conclusion, our simulations suggest that the complex GRN that underlies the attainment of VM, IM, and FM cell identities also restricts, to a large extent, the temporal pattern of transitions among them as found for the floral organ specification GRN reported by Álvarez-Buylla et al. (2008).

Interestingly, in the case of gain of function simulations of *XAL2*, *SOC1*, and *AGL24*, the same pattern of temporal transitions as in wild-type was recovered, but in these cases the maximum probability of the mixed IM/FM identity occurs after the IM and before the FM configurations (Figure 7B–7D). This analysis also recovers all the possible transitions among the steady states (Figure 7E and 7F). The net transition rate was positive for the IM to FM direction in all the lines tested, but was lower under gain of function lines in comparison with the wild-type (Supplemental Figure 4). This means that the net probability flow preferentially follows the direction from IM to FM, both in wild-type and in each of the overexpression lines of *XAL2*, *SOC1*, and *AGL24* (Figure 7F). These results are consistent with the observed most probable temporal order of transitions in plants. Likewise, the results do not support the hypothesis of an induced, reverse

	AP1	LFY	SOC1	AGL24	XAL2	TFL1	FT	GA	CO
VM	0	0	0	0	0	1	0	[a]	0
IM	0	0	[a]	[a]	[a]	1	[a]	[a]	[a]
FM	1	1	[a]	[a]	[a]	0	[a]	[a]	[a]

Table 1. Observed Expression States of the Genes Considered in the Network Model in Wild-Type Plants during Different Stages of the SAM Development.

[a]Any possible value of the node in the network.

rate of transition from FM to IM or IM/FM cells as an explanation of the observed phenotype in the overexpressors, as both reverse transitions (FM to IM and FM to IM/FM) showed a negative net transition rate (Figure 7E). Overall, the results of the stochastic EL analysis suggest that instead of an accelerated rate of transition in the forward (IM to FM) direction, it is the novel potentiality of the IM state to now choose between two preferential (positive net transition rate) fate decisions (FM or IM/FM phenotypes) induced by gene overexpression that accounts for the observed promiscuous IM/FM state in such flowers (Figure 7F and Supplemental Figure 4).

DISCUSSION

In this work we have shown, in contrast to previous expectations (Schönrock et al., 2006; Garay-Arroyo et al., 2013), that *XAL2* is expressed in the IM and FM and is a key player in the complex GRN underlying SAM transitions (Figure 6). *XAL2* is a promoter of flowering in response to multiple signals and is also important for FM maintenance and determinacy. We propose a GRN and EL modeling approach that together provides a mechanistic dynamic framework to explain the role of *XAL2* at the SAM and the apparently paradoxical phenotypes of its overexpression. Moreover, such a modeling framework constitutes a systemic mechanistic explanation for the observed patterns of expression of multiple genes underlying VM, IM, and FM cell fates, and the observed transitions among them in wild-type *Arabidopsis*. It thus constitutes a useful framework to incorporate additional components and interactions that participate in SAM development. Finally, it provides an explanation for *AGL24*, *SOC1*, and their homolog overexpression phenotypes in *Arabidopsis* (Borner et al., 2000; Michaels et al., 2003; Ferrario et al., 2004; Masiero et al., 2004; Yu et al., 2004; Liu et al., 2007; Trevaskis et al., 2007; Fornara et al., 2008).

XAL2 Promotes Flowering Transition

XAL2 participates in flowering transition in response to more than one signal, having a higher impact under non-inductive photoperiod conditions (Figure 1C). Flowering time is not so clearly affected in the *xal2* alleles, under all conditions tested, as is *soc1*, probably because SOC1 and AGL24 are able to directly activate *LFY* independently of *XAL2* (Lee et al., 2008; Liu et al., 2008). We proved that CO positively regulates *XAL2* and that the latter positively regulates *SOC1* and probably *AGL24* (Figure 2C–2E). Being *soc1* epistatic over *xal2* under LD condition confirms this result (Figure 1B). We also proved that under SD condition, and in response to GA, *xal2* is affected in bolting time and *xal2-2 soc1-6* has an additive effect compared with the parental plants (Figure 1C and 1D and Supplemental Table 1). These results could imply that they act independently over *LFY* and *AP1* regulation, or that they are part of the same regulatory module. We argue that *XAL2* is probably part of the same GRN in which *SOC1* participates, integrating at least some of the flowering transition pathways in response to different signals. In fact, the spatial and temporal patterns of expression of *XAL2*, and its loss and gain of function phenotypes, resemble those corresponding to *SOC1* and *AGL24* lines (Borner et al., 2000; Yu et al., 2004; Liu et al., 2007; Gregis et al., 2009), thus suggesting that *XAL2* is part of the *SOC1–AGL24* regulatory module. Moreover, XAL2 interacts with SOC1 and AGL24 according to yeast two-hybrid data (de Folter et al., 2005; Immink et al., 2009).

XAL2 Overexpression Affects FM Maintenance and Determinacy by Up-Regulating *TFL1* and *WUS*

After the flowering transition, LFY, AP1, and CAL are necessary for FM identity (Weigel et al., 1992; Bowman et al., 1993; Ferrándiz et al., 2000) by repressing the IM genes, particularly *TFL1* (Shannon and Meeks-Wagner, 1991; Schultz and Haughn, 1993; Gustafson-Brown et al., 1994; Mandel and Yanofsky, 1995; Liljegren et al., 1999). During the first and second stages of FM development, SOC1, AGL24, and SVP maintain FM identity in collaboration with AP1 by repressing *AG* and *SEP3* (Gregis et al., 2006, 2009; Liu et al., 2009). At stage 3 of flower development, LFY and AP1 repress the expression of the "flowering genes," allowing the transcription of the floral organ identity genes (Yu et al., 2004; Liu et al., 2007). LFY and WUS, among other genes, induce the expression of *AG* during this stage, which in turn represses *WUS* at stage 6, together with other proteins (Lenhard et al., 2001; Lohmann, et al., 2001; Gómez-Mena et al., 2005; Lee et al., 2005; Sun et al., 2009; Sun and Ito, 2010; Liu et al., 2011). This event drastically affects the FM stem cells, which stop proliferating (Mizukami and Ma, 1997).

These experimental data indicate that certain genes have clear effects in the FM when their expression is depleted or augmented; however, we think that FM identity, maintenance, and determinacy emerge from a complex GRN in which spatio-temporal regulations of *SOC1*, *AGL24*, *SVP*, and *XAL2* are also important. Indeed, in this study we have shown that overexpression of *XAL2* affects FM maintenance and yields phenotypes similar to those reported for the overexpression lines of *SOC1*, *AGL24*, and their homologs (Borner et al., 2000; Michaels et al., 2003; Ferrario et al., 2004; Masiero et al., 2004; Yu et al., 2004; Liu et al., 2007; Trevaskis et al., 2007; Fornara et al., 2008). More importantly, we demonstrate that overexpression of any of these genes is sufficient to induce *TFL1* expression (Figure 4A and 4C), suggesting that mis-regulation of *TFL1* underlies the "leaf-like" flower phenotype observed in the overexpression of these three MADS-box genes. In this regard, Hanano and

Goto (2011) had demonstrated that TFL1 acts as a transcriptional repressor, and the *35S::TFL1-SRDX* line phenotype reported by these authors is, in fact, very similar to the *XAL2* phenotype reported here (Figure 3).

Overexpression of *SOC1*, *AGL24*, or *XAL2* genes affects SAM transitions, causing premature flowering and *LFY/AP1* up-regulation (Figures 2, 4, and 5). At the same time, we have proved that overexpression of these MADS-box genes induces higher levels of *TFL1* mRNA accumulation compared with wild-type plants (Figure 4A and 4C). Furthermore, we have shown that XAL2 directly binds to *TFL1* regulatory sequences using the overexpression line *35S::GFP:XAL2* (Figure 4B). Interestingly, one of these binding sites (fragment V of the *TFL1* 3′ region amplified in our ChIP assay) corresponds to one of the binding sites of AP1, which has been demonstrated to be important for direct repression of *TFL1* (Kaufman et al., 2010). More recently, it was demonstrated that SOC1, AGL24, SVP, and SEP4 cooperate with AP1 in this action (Liu et al., 2013). However, it is possible that, when overexpressed, higher ratios of XAL2, SOC1, or AGL24 over AP1 are able to compete for the same binding site, affecting *TFL1* transcription in an opposite way. The partial complementation of the vegetative and indeterminacy features of the *35S::XAL2* line by crossing it with *35S::AP1* supports this hypothesis (Figure 5). If *TFL1* and, probably, other genes important for IM identity are ectopically expressed in the FM, this would explain the inflorescence characteristics of those flowers even in the presence of *AP1* which is not down-regulated (Figures 2 and 4; Supplemental Figure 3), and probably not mis-localized either, as reported for *AGL24/SVP* homolog *OsMADS47* overexpression line (Fornara et al., 2008). In this sense, the FM does not change its identity through a floral reversion process. Instead it behaves differently, probably having a mixed IM/FM identity, due to an altered behavior of the GRN (Figures 6 and 7).

Heterochronic ''floral reversion'' has been shown to be dependent on light and gibberellin signaling that affects a signal coming from the leaves to the SAM (Okamuro et al., 1996; Hempel et al., 2000). We now know that this signal is FT (Jaeger and Wigge, 2007; Müller-Xing et al., 2014). During flowering transition, this protein competes with TFL1 for FD, and this association up-regulates *SOC1* and *AP1* in the anlagen (Abe et al., 2005; Wigge et al., 2005; Hanano and Goto, 2011; Jaeger et al., 2013). In the overexpression lines of *XAL2*, *SOC1*, and *AGL24*, up-regulation of *TFL1* or delayed expression of *FT* under SD condition would affect such balance until endogenous FT protein attains certain levels during *Arabidopsis* inflorescence development. This would explain why the acropetal flowers show a wild-type phenotype while the early ones show IM features. This and related hypotheses could be tested by expanding the dynamic GRN and EL modeling framework proposed here.

Early Flowering and FM Phenotypes of the *XAL2* Overexpression Line under LD Condition are Reconciled Using a GRN Model and EL Analysis

We proposed GRN and EL models that provide a framework for mechanistic explanations of SAM transitions in wild-type plants, but also the complex loss and gain of function phenotypes of *XAL2* and other regulators of SAM transitions. In particular, this provides a novel framework with which to evaluate floral reversion. Floral reversion has been defined as the reappearance of vegetative traits during flower development or the loss of FM determinacy after floral organs are formed. This uncommon process in *Arabidopsis* has been attributed to reversion of the FM to the IM identity, particularly in the *lfy* and *ap1* mutants (Battey and Lyndon, 1990; Okamuro et al., 1993, 1996, 1997; Tooke et al., 2005). In contrast, based on previous data and the experimental results summarized here, we postulate an alternative explanation for the so-called floral reversion in the case of the *SOC1*, *AGL24*, and *XAL2* overexpression lines.

Our results of the deterministic GRN model suggest that a mixed meristem identity is attained as a steady state when *XAL2*, *AGL24*, or *SOC1* are overexpressed, while the same GRN yields normal configurations when the same genes are kept to ''0.'' Indeed, based on our experimental data, in the model the IM/FM identity is the result of the positive regulation of these three MADS-box genes over *TFL1*, *LFY*, and *AP1*. When either of these genes is overexpressed, *TFL1* and *AP1* or *LFY* are activated, while at the same time the multiple feedback loops among them stabilize their expression, thus yielding the IM/FM identity (Figures 6 and 7).

The EL simulations suggest a mechanism by which *35S::SOC1*, *35S::AGL24*, or *35S::XAL2* cause a fraction of the cell population at the IM to acquire a mixed IM/FM identity (Figure 7). This could be explained by two alternative hypotheses. During normal developmental VM > IM > FM transitions, a fraction of cells may attain the new mixed identity IM/FM. Under this circumstance the establishment of the antagonistic relationship between IM and FM regulators may be weakened. On the other hand, an induced, reverse rate of transition from FM to IM or IM/FM cells could account for the results. The modeling results show that the first one is the most probable one, and the overexpressor global transition pattern is: VM > IM > IM/FM > FM (Figure 7F and Supplemental Figure 4). Therefore, for this and similar cases the term ''floral reversion'' should be avoided.

Loss of FM Determinacy in *XAL2* Overexpression Lines under SD Condition

Constitutive expression of *XAL2* also affects floral determinacy under SD condition. Here we showed that under this condition new inflorescences develop from the inside of the carpels of the basipetal flowers (Figure 3H, 3J, and 3K). At the molecular level, this may be explained in two different ways: either the presence of XAL2 prevents *WUS* repression or ectopic expression of this gene is sufficient to up-regulate *WUS*. We observed that *WUS* expression in the *35S::XAL2* is maintained after stage 6, enabling stem cells to remain active (Figures 3J, 3K, and 4F). At this point, we cannot know if the FM maintenance and indeterminacy phenotypes observed in the overexpression lines of *XAL2*, *SOC1*, or *AGL24* are due to a dominant negative effect or to gain of function. Interestingly, overexpression of *XAL2* or *SOC1* represses each other (Figure 2C and 2F), indicating that in these lines altered protein complexes could be formed. These hypotheses can be tested using an expanded GRN module including additional SAM genes. Furthermore, such a model could address whether FM

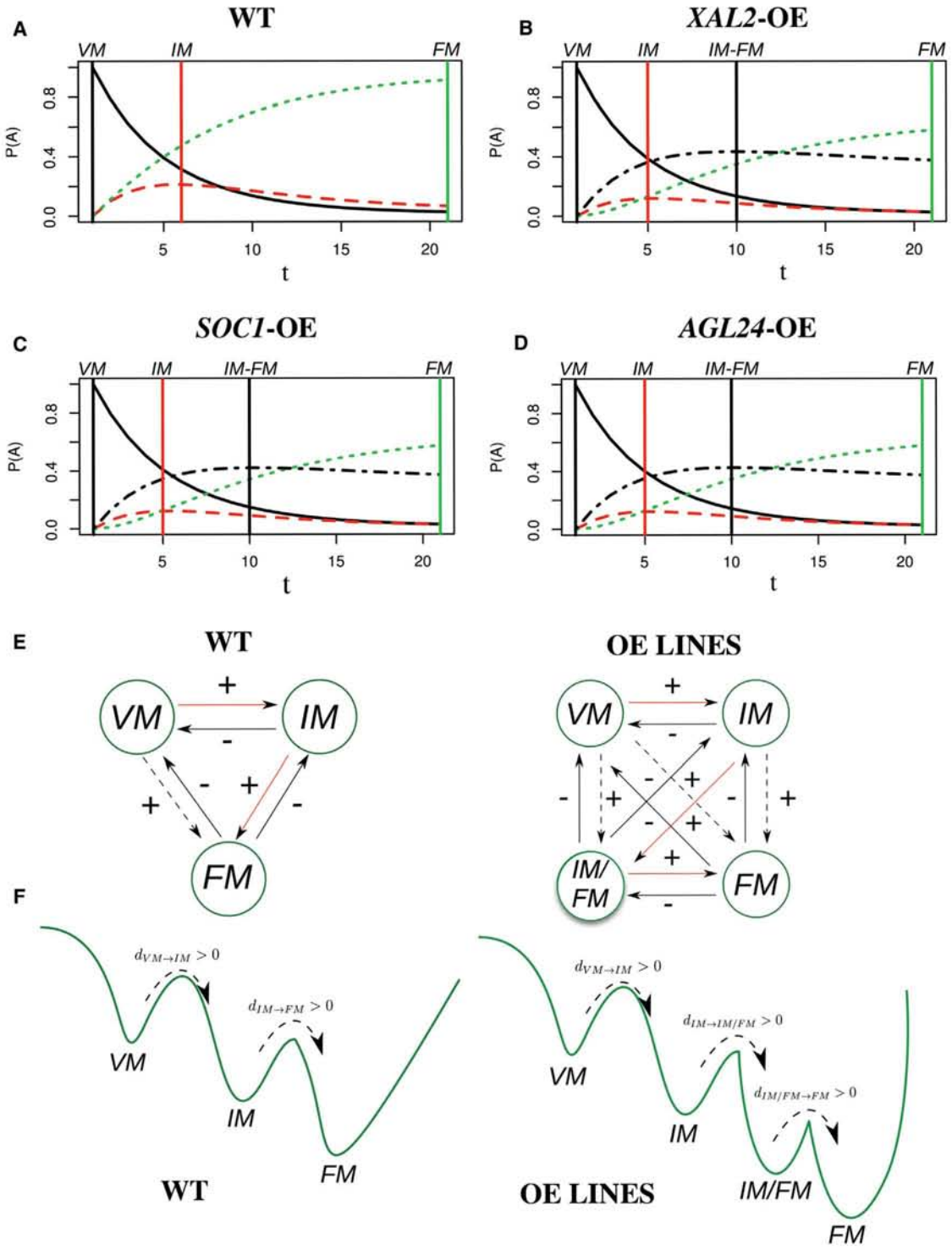


Figure 7. Epigenetic Landscape Analysis for the *XAL2* Regulatory Network Module.

(A–D) Temporal sequence of cell-fate attainment pattern under the stochastic GRN model during SAM cell-fate transitions. The maximum probability *p* of attaining each attractor, as a function of time (in iteration steps) is shown for **(A)** WT, **(B)** *XAL2* overexpression (*XAL2-OE*), **(C)** *SOC1-OE*, and **(D)** *AGL24-OE*. Vertical lines mark the time at which maximum probability of each steady state (i.e., cell fate, VM, IM, FM, or IM/FM) is attained. Note that the maximum probability for each steady state is 1. The most probable sequence of cell-fate attainment for the WT is VM, IM, FM; and for OE lines VM, IM (IM/FM), FM. The value of the error probability used in this case was $\xi = 0.05$. The same patterns were obtained with error probabilities from 0.01 to 0.1 (data not shown).

(E) Schematic representation of the possible transitions between pairs of steady states (cell fates at the SAM) for WT and OE lines. Arrows indicate the directionality of the transitions. Above each arrow a sign (+) or (−) indicates whether the calculated net transition rate between the corresponding

(legend continued on next page)

to IM transition in the indeterminate carpels, which corresponds to cell reprogramming, is favored under *XAL2* or other MADS-box overexpression.

METHODS

Plant Material and Selection of Mutant Lines

Arabidopsis thaliana wild-type and mutant plants used in this study were Col-0 with the exception of *ap1-1 cal-5* and *tfl1-2*, which are in L*er* ecotype. Mutant alleles *xal2-1* and *xal2* were described previously (Garay-Arroyo et al., 2013). The *soc1-6* (SALK_138131; Wang et al., 2009), *ful-7* (SALK_033647; Wang et al., 2009), and *agl24-4* (GK674F05.03/N385337) mutant seeds were provided by the Arabidopsis Biological Resource Center or the Nottingham Arabidopsis Stock Centre, and the homozygous alleles were selected using the primers shown in Supplemental Table 4.

Plant Growth Conditions and Flowering Time Measurements

Seedlings were grown on vertical plates with 0.2× Murashige and Skoog (MS) medium (Murashige and Skoog, 1962) containing 1% sucrose. For flowering experiments, plants were grown on soil (Metromix 200) under LD (16 h light/8 h dark) or SD (8 h light/16 h dark) condition at 22°C. For GA_3 treatment, plants were grown under SD condition for 2 weeks before they were sprayed with 100 μM GA_3 twice a week until flowering. For vernalization experiments, seeds were plated on MS medium and kept in the dark for 8 weeks at 4°C and then transferred to soil and grown under LD condition. Flowering transition was measured as bolting time (days after seed sowing required for the stem to grow to 1 cm long) and by the RLN at bolting. Inflorescences for *in situ* hybridization were collected when the stem reached 10 cm long. These comprised FM at different developmental stages.

Plasmid Constructs and Plant Selection

The *XAL2* gene was amplified from Col-0, using the XAL2g-F 5′-AGAAGAATGGTGAGGGGAAA-3′ and XAL2g-R 5′-ATGTTAGTTTGAAGGAGGAA-3′ primers. The 3603 nt DNA fragment was cloned in the pCR8/GW/TOPO-TA vector, and verified by sequencing. It was then recombined into either overexpression vectors: pGD625 (de Folter et al., 2006) or the pK7WGF2 that includes GFP (Karimi et al., 2002) carrying a kanamycin and spectinomycin/streptomycin resistance cassette, respectively. Kanamycin (50 μg/ml) resistant plants were selected on plates.

In Situ Hybridization Analysis

In situ hybridization was performed according to Tapia-López et al. (2008). *In vitro* transcription with the DIG RNA labeling Kit (Roche Molecular Biochemicals) was performed to generate the antisense *XAL2* probe using as a template the XAL2-F 5′-GTTTCCTCCTTCAAACTAACA-3′ and XAL2-R 5′-GCAACTGCTAAATTCAGTAAG-3′ amplified cDNA fragment cloned into p-GEM-T.

Quantitative Real-Time RT–PCR

Aerial tissue from three independent biological replicates (15 plants each) was used for total RNA extraction with Trizol reagent, and two independent cDNAs were reverse transcribed using Superscript II (Invitrogen). We amplified *PDF2* (AT1G13320) and *UPL7* (AT1G13320) as positive internal controls (Czechowski et al., 2005), and their stability across the compared samples was confirmed using geNorm (Vandesompele et al., 2002). Amplification efficiencies were analyzed using Real Time PCR Miner (Zhao and Fernald, 2005), and relative expression was calculated using the ΔΔCT method (Vandesompele et al., 2002). Primer sequences are presented in Supplemental Table 4.

Microscopy

An Olympus SZ60 dissecting microscope with C-5060 digital camera was used for light microscopy. Sectioned carpels were fixed in 4% paraformaldehyde, dehydrated in ethanol series, and embedded in paraffin. Sections (8 μm) were stained with toluidine blue 0.05%. For scanning electron microscopy, plant material was fixed at 4°C overnight in 50% ethanol, 5% acetic acid, and 3.7% formaldehyde in 0.025 M phosphate buffer (pH 7.0). Samples were subsequently washed twice (30 min) in 70% ethanol in the same phosphate buffer, followed by 0.05 M phosphate buffer (pH 7.0). Samples were dehydrated gradually to ethanol 100%, and dried in liquid carbon dioxide at the critical point. Finally, samples were covered with gold using a sputter coater and observed with a scanning electron microscope.

TFL1 ChIP Assays

Wild-type and the *35S::GFP-XAL2* line were grown in MS plates under LD condition and inflorescence tissue (0.5 g) was fixed for 20 min. Chromatin was solubilized with a sonicator by three pulses of 15 s each. Immunoprecipitation was performed overnight using anti-GFP rabbit IgG fraction (A11122; Invitrogen) and protein A agarose beads (Santa Cruz). Samples were treated with proteinase K after elution followed by precipitation. Template ChIP DNA was diluted and amplified for 35–40 cycles (de Folter et al., 2007; de Folter, 2011). Primer pairs were designed in flanking regions of CArG boxes found along 2 kb upstream of the start codon, as well as 4.6 kb downstream of the *TFL1* gene (Supplemental Table 4).

GRN Model: Recovery of Gene Expression Profiles Characteristic of VM, IM, and FM Cell Types

The GRN was modeled using a discrete multi-state GRN formalism as described by Espinosa-Soto et al. (2004) and Álvarez-Buylla et al. (2010a, 2010b).

Stochastic GRN Model Implementation: EL Approach

To explore the patterns of cell-fate attainment and transition among cells, a discrete stochastic GRN dynamic model was implemented as an extension of the deterministic Boolean model described in the previous section. Stochasticity is modeled by introducing a constant probability of error for the deterministic Boolean logical functions according to:

$$x_i(t+1) = \begin{cases} f_i(t), & \text{with prob } 1-\xi \\ 1-f_i(t), & \text{with prob } \xi \end{cases}.$$

We followed Álvarez-Buylla et al. (2008). This approach yields a probability matrix that was then used to describe how the probability of being in a particular steady state changes in time by iterating the dynamic equation

$$p_x(t+1) = p_x(t)P,$$

where P is the transition probability matrix and $p_x(t)$ the distribution vector specifying the proportion of cells or the probability of a single one being in each steady state at a given time.

attractors is positive or negative. Red arrows represent the globally consistent ordering for the 3(4) attractors: the order of the attractors in which all individual transition has a positive net rate, resulting in a global probability flow across the EL as also shown in **(F)** (see Supplemental Methods).
(F) Schematic representation of the EL of the GRN modeled here. The relative barrier heights represent the hierarchy of calculated positive net probability rates, which altogether determine a consistent global ordering of the relative steady-state stabilities. According to the net probability rates, only one set of ordered transition (VM > IM > [IM/FM] > FM) produces a positive probability flow (see Supplemental Methods). As a result, a global developmental gradient in the EL is produced. Importantly this 2D representation is for illustrative purposes only and, as such, does not represent scales based on exact calculated values.

EL Exploration

To explore the EL associated with a GRN, the number, depth, width, and relative position of the GRN attractors are represented by the hills and valleys of Waddington's (EL) metaphor (Álvarez-Buylla et al., 2008). In addition to the calculation of the most probable temporal cell-fate pattern, a discrete stochastic GRN model allows calculations of the shortest and fastest pathways of cell-fate transitions, as well as possible restrictions of some cell-fate transitions that also emerge from the GRN topology and the associated EL. We calculated the mean first passage time (MFPT) between each pair of possible transitions to uncover which of these is more feasible. MFPT was estimated numerically by using the transition probabilities among steady states from a large number of samples of paths simulated as a finite Markov chain process (Wilkinson, 2011). The MFPT from one steady state (i) to another (j) corresponds to the average value of the number of steps taken to visit attractor j for the first time, given that the entire probability mass was initially localized at steady state i. The average is taken over a large number of realizations (simulations). Based on the MFPT values, a net transition rate between steady states i and j can be defined as follows: $d_{i\to j} = 1/\mathrm{MFPT}_{i\to j} - 1/\mathrm{MFPT}_{j\to i}$. This quantity effectively measures the facility by which a state transits from one state to another as a net probability flow (Zhou et al., 2014). For all stochastic modeling, robustness was assessed by comparing three different values for the error probability (0.01, 0.05, 0.1). The number of simulated samples was increased until stable results were attained. See also Supplemental Methods.

SUPPLEMENTAL INFORMATION

Supplemental Information is available at *Molecular Plant Online*.

FUNDING

This research was supported by CONACyT (81433; 180098; 180380; 167705; 152649; 147675; 177739), PAPIIT, UNAM (IN204011-3; IN203214-3; IN203113-3; IN203814-3), and UC-MEXUS ECO-IE415 grants. E.R.A.B. was supported by the Miller Institute for Basic Research in Science, University of California, Berkeley, USA.

ACKNOWLEDGMENTS

This paper constitutes a partial fulfillment of the graduate program "Doctorado en Ciencias Biomédicas of the Universidad Nacional Autónoma de México" in which Rigoberto V. Pérez-Ruiz developed this project. We acknowledge Dr. Yanofsky and Dr. Pelaz for helping at early stages of this work. We thank researchers who shared their lines: Dr. Alonso provided *tfl1-1* mutant; Dr. Yu the *co-1* allele, and the *35S::AP1* and *35S::AGL24* lines; Dr. Yanofsky the *lfy-9*, *ap1-15*, and *ap1-1 cal-5* mutants, and Dr. Lee the *agl20-101D* line. Diana Romo, Dr. Martínez-Silva, and K. González-Aguilera helped with logistical and technical tasks, and Dr. Espinosa-Matias SEM preparations (Facultad de Ciencias, UNAM). We thank Rich Jorgensen for editing. No conflict of interest declared.

Received: December 10, 2014
Revised: December 10, 2014
Accepted: January 5, 2015
Published: January 28, 2015

REFERENCES

Abe, M., Kobayashi, Y., Yamamoto, S., Daimon, Y., Yamaguchi, A., Ikeda, Y., Ichinoki, H., Notaguchi, M., Goto, K., and Araki, T. (2005). FD, a bZIP protein mediating signals from the floral pathway integrator FT at the shoot apex. Science **309**:1052–1056.

Alvarez, J., Guli, C.L., Yu, X.H., and Smyth, D.R. (1992). Terminal flower: a gene affecting inflorescence development in *Arabidopsis thaliana*. Plant J. **2**:103–116.

Álvarez-Buylla, E.R., Pelaz, S., Liljegren, S.J., Gold, S.E., Burgeff, C., Ditta, G.S., Ribas de Pouplana, L., Martinez-Castilla, L., and Yanofsky, M.F. (2000). An ancestral MADS-box gene duplication occurred before the divergence of plants and animals. Proc. Natl. Acad. Sci. USA **97**:5328–5333.

Álvarez-Buylla, E.R., Balleza, E., Benitez, M., Espinosa-Soto, C., and Padilla-Longoria, P. (2008). Gene regulatory network models: a dynamic and integrative approach to development. SEB Exp. Biol. Ser. **61**:113–139.

Álvarez-Buylla, E.R., Benitez, M., Corvera-Poire, A., Chaos Cador, A., de Folter, S., Gamboa de Buen, A., Garay-Arroyo, A., Garcia-Ponce, B., Jaimes-Miranda, F., Perez-Ruiz, R.V., et al. (2010a). Flower development. *Arabidopsis* Book **8**:e0127.

Álvarez-Buylla, E.R., Azpeitia, E., Barrio, R., Benitez, M., and Padilla-Longoria, P. (2010b). From ABC genes to regulatory networks, epigenetic landscapes and flower morphogenesis: making biological sense of theoretical approaches. Semin. Cell Dev. Biol. **21**:108–117.

An, H., Roussot, C., Suarez-Lopez, P., Corbesier, L., Vincent, C., Pineiro, M., Hepworth, S., Mouradov, A., Justin, S., Turnbull, C., et al. (2004). CONSTANS acts in the phloem to regulate a systemic signal that induces photoperiodic flowering of *Arabidopsis*. Development **131**:3615–3626.

Andrés, F., and Coupland, G. (2012). The genetic basis of flowering responses to seasonal cues. Nat. Rev. Genet. **13**:627–639.

Balasubramanian, S., Sureshkumar, S., Lempe, J., and Weigel, D. (2006). Potent induction of *Arabidopsis thaliana* flowering by elevated growth temperature. PLoS Genet. **2**:e106.

Battey, N.H., and Lyndon, R.F. (1990). Reversion of flowering. Bot. Rev. **56**:162–189.

Benlloch, R., Berbel, A., Serrano-Mislata, A., and Madueno, F. (2007). Floral initiation and inflorescence architecture: a comparative view. Ann. Bot. **100**:1609.

Blázquez, M.A., and Weigel, D. (2000). Integration of floral inductive signals in *Arabidopsis*. Nature **404**:889–892.

Blázquez, M.A., Green, R., Nilsson, O., Sussman, M.R., and Weigel, D. (1998). Gibberellins promote flowering of *Arabidopsis* by activating the LEAFY promoter. Plant Cell **10**:791–800.

Blázquez, M.A., Ahn, J.H., and Weigel, D. (2003). A thermosensory pathway controlling flowering time in *Arabidopsis thaliana*. Nat. Genet. **33**:168–171.

Borner, R., Kampmann, G., Chandler, J., Gleissner, R., Wisman, E., Apel, K., and Melzer, S. (2000). A MADS domain gene involved in the transition to flowering in *Arabidopsis*. Plant J. **24**:591–599.

Bowman, J.L., Alvarez, J., Weigel, D., Meyerowitz, E.M., and Smyth, D.R. (1993). Control of flower development in *Arabidopsisthaliana* by *APETALA*1 and interacting genes. Development **119**: 721–743.

Bradley, D., Ratcliffe, O., Vincent, C., Carpenter, R., and Coen, E. (1997). Inflorescence commitment and architecture in *Arabidopsis*. Science **275**:80–83.

Chen, L.J., Cheng, J.C., Castle, L., and Sung, Z.R. (1997). EMF genes regulate *Arabidopsis* inflorescence development. Plant Cell **9**:2011–2024.

Conti, L., and Bradley, D. (2007). TERMINAL FLOWER1 is a mobile signal controlling *Arabidopsis* architecture. Plant Cell **19**:767–778.

Czechowski, T., Stitt, M., Altmann, T., Udvardi, M.K., and Scheible, W.R. (2005). Genome-wide identification and testing of superior reference genes for transcript normalization in *Arabidopsis*. Plant Physiol. **139**:5–17.

de Folter, S. (2011). Protein tagging for chromatin immunoprecipitation from *Arabidopsis*. Methods Mol. Biol. **678**:199–210.

de Folter, S., Immink, R.G.H., Kieffer, M., Parenicová, L., Henz, S.R., Weigel, D., Busscher, M., Kooiker, M., Colombo, L., Kater, M.M.,

et al. (2005). Comprehensive interaction map of the *Arabidopsis* MADS box transcription factors. Plant Cell **17**:1424–1433.

de Folter, S., Shchennikova, A.V., Franken, J., Busscher, M., Baskar, R., Grossniklaus, U., Angenent, G.C., and Immink, R.G.H. (2006). A B-sister MADS-box gene involved in ovule and seed development in petunia and *Arabidopsis*. Plant J. **47**:934–946.

de Folter, S., Urbanus, S.L., van Zuijlen, L.G.C., Kaufmann, K., and Angenent, G.C. (2007). Tagging of MADS domain proteins for chromatin immunoprecipitation. BMC Plant Biol. **7**:47.

Dorca-Fornell, C., Gregis, V., Grandi, V., Coupland, G., Colombo, L., and Kater, M.M. (2011). The *Arabidopsis* SOC1-like genes AGL42, AGL71 and AGL72 promote flowering in the shoot apical and axillary meristems. Plant J. **67**:1006–1017.

Espinosa-Soto, C., Padilla-Longoria, P., and Álvarez-Buylla, E.R. (2004). A gene regulatory network model for cell-fate determination during *Arabidopsis thaliana* flower development that is robust and recovers experimental gene expression profiles. Plant Cell **16**:2923–2939.

Ferrándiz, C., Gu, Q., Martienssen, R., and Yanofsky, M.F. (2000). Redundant regulation of meristem identity and plant architecture by FRUITFULL, APETALA1 and CAULIFLOWER. Development **127**: 725–734.

Ferrario, S., Busscher, J., Franken, J., Gerats, T., Vandenbussche, M., Angenent, G.C., and Immink, R.G.H. (2004). Ectopic expression of the petunia MADS box gene UNSHAVEN accelerates flowering and confers leaf-like characteristics to floral organs in a dominant-negative manner. Plant Cell **16**:1490–1505.

Fornara, F., Gregis, V., Pelucchi, N., Colombo, L., and Kater, M. (2008). The rice StMADS11-like genes OsMADS22 and OsMADS47 cause floral reversions in *Arabidopsis* without complementing the svp and agl24 mutants. J. Exp. Bot. **59**:2181–2190.

Garay-Arroyo, A., Ortiz-Moreno, E., de la Paz Sánchez, M., Murphy, A.S., García-Ponce, B., Marsch-Martínez, N., de Folter, S., Corvera-Poire, A., Jaimes-Miranda, F., Pacheco-Escobedo, M.A., et al. (2013). The MADS transcription factor XAL2/AGL14 modulates auxin transport during *Arabidopsis* root development by regulating PIN expression. EMBO J. **32**:2884–2895.

Gómez-Mena, C., de Folter, S., Costa, M.M., Angenent, G.C., and Sablowski, R. (2005). Transcriptional program controlled by the floral homeotic gene AGAMOUS during early organogenesis. Development **132**:429–438.

Gramzow, L., Ritz, M.S., and Theissen, G. (2010). On the origin of MADS-domain transcription factors. Trends Genet. **26**:149–153.

Gregis, V., Sessa, A., Colombo, L., and Kater, M.M. (2006). AGL24, SHORT VEGETATIVE PHASE, and APETALA1 redundantly control AGAMOUS during early stages of flower development in *Arabidopsis*. Plant Cell **18**:1373–1382.

Gregis, V., Sessa, A., Dorca-Fornell, C., and Kater, M.M. (2009). The *Arabidopsis* floral meristem identity genes AP1, AGL24 and SVP directly repress class B and C floral homeotic genes. Plant J. **60**:626–637.

Gustafson-Brown, C., Savidge, B., and Yanofsky, M.F. (1994). Regulation of the *Arabidopsis* floral homeotic gene APETALA1. Cell **76**:131–143.

Halliday, K.J., Salter, M.G., Thingnaes, E., and Whitelam, G.C. (2003). Phytochrome control of flowering is temperature sensitive and correlates with expression of the floral integrator FT. Plant J. **33**:875–885.

Han, P., Garcia-Ponce, B., Fonseca-Salazar, G., Álvarez-Buylla, E.R., and Yu, H. (2008). AGAMOUS-LIKE 17, a novel flowering promoter, acts in a FT-independent photoperiod pathway. Plant J. **55**:253–265.

Hanano, S., and Goto, K. (2011). *Arabidopsis* TERMINAL FLOWER1 is involved in the regulation of flowering time and inflorescence development through transcriptional repression. Plant Cell **23**:3172–3184.

Hartmann, U., Hohmann, S., Nettesheim, K., Wisman, E., Saedler, H., and Huijser, P. (2000). Molecular cloning of SVP: a negative regulator of the floral transition in *Arabidopsis*. Plant J. **21**:351–360.

Hempel, F.D., Welch, D.R., and Feldman, L.J. (2000). Floral induction and determination: where is flowering controlled? Trends Plant Sci. **5**:17–21.

Huala, E., and Sussex, I.M. (1992). LEAFY interacts with floral homeotic genes to regulate *Arabidopsis* floral development. Plant Cell **4**: 901–913.

Immink, R.G.H., Tonaco, I.A.N., de Folter, S., Shchennikova, A., van Dijk, A.D.J., Busscher-Lange, J., Borst, J.W., and Angenent, G.C. (2009). SEPALLATA3: the 'glue' for MADS box transcription factor complex formation. Genome Biol. **10**:R24.

Irish, V.F., and Sussex, I.M. (1990). Function of the apetala-1 gene during *Arabidopsis* floral development. Plant Cell **2**:741–753.

Jaeger, K.E., and Wigge, P.A. (2007). FT protein acts as a long-range signal in *Arabidopsis*. Curr. Biol. **17**:1050–1054.

Jaeger, K.E., Pullen, N., Lamzin, S., Morris, R.J., and Wigge, P.A. (2013). Interlocking feedback loops govern the dynamic behavior of the floral transition in *Arabidopsis*. Plant Cell **25**:820–833.

Karimi, M., Inze, D., and Depicker, A. (2002). GATEWAY vectors for *Agrobacterium*-mediated plant transformation. Trends Plant Sci. **7**:193–195.

Kaufmann, K., Wellmer, F., Muino, J.M., Ferrier, T., Wuest, S.E., Kumar, V., Serrano-Mislata, A., Madueno, F., Krajewski, P., Meyerowitz, E.M., et al. (2010). Orchestration of floral initiation by APETALA1. Science **328**:85–89.

Kaufmann, K., Nagasaki, M., and Jauregui, R. (2011). Modelling the molecular interactions in the flower developmental network of *Arabidopsis thaliana*. Stud. Health Technol. Inform. **162**:279–297.

Koornneef, M., Hanhart, C.J., and van der Veen, J.H. (1991). A genetic and physiological analysis of late flowering mutants in *Arabidopsis thaliana*. Mol. Gen. Genet. **229**:57–66.

Lee, J., and Lee, I. (2010). Regulation and function of SOC1, a flowering pathway integrator. J. Exp. Bot. **61**:2247–2254.

Lee, H., Suh, S.S., Park, E., Cho, E., Ahn, J.H., Kim, S.G., Lee, J.S., Kwon, Y.M., and Lee, I. (2000). The AGAMOUS-LIKE 20 MADS domain protein integrates floral inductive pathways in *Arabidopsis*. Genes Dev. **14**:2366–2376.

Lee, J.Y., Baum, S.F., Alvarez, J., Patel, A., Chitwood, D.H., and Bowman, J.L. (2005). Activation of CRABS CLAW in the nectaries and carpels of *Arabidopsis*. Plant Cell **17**:25–36.

Lee, J.H., Yoo, S.J., Park, S.H., Hwang, I., Lee, J.S., and Ahn, J.H. (2007). Role of SVP in the control of flowering time by ambient temperature in *Arabidopsis*. Genes Dev. **21**:397–402.

Lee, J., Oh, M., Park, H., and Lee, I. (2008). SOC1 translocated to the nucleus by interaction with AGL24 directly regulates LEAFY. Plant J. **55**:832–843.

Lenhard, M., Bohnert, A., Jurgens, G., and Laux, T. (2001). Termination of stem cell maintenance in *Arabidopsis* floral meristems by interactions between WUSCHEL and AGAMOUS. Cell **105**:805–814.

Li, D., Liu, C., Shen, L., Wu, Y., Chen, H., Robertson, M., Helliwell, C.A., Ito, T., Meyerowitz, E., and Yu, H. (2008). A repressor complex governs the integration of flowering signals in *Arabidopsis*. Dev. Cell **15**:110–120.

Liljegren, S.J., Gustafson-Brown, C., Pinyopich, A., Ditta, G.S., and Yanofsky, M.F. (1999). Interactions among APETALA1, LEAFY, and TERMINAL FLOWER1 specify meristem fate. Plant Cell **11**:1007–1018.

Liu, C., Zhou, J., Bracha-Drori, K., Yalovsky, S., Ito, T., and Yu, H. (2007). Specification of *Arabidopsis* floral meristem identity by repression of flowering time genes. Development **134**:1901–1910.

Liu, C., Chen, H., Er, H.L., Soo, H.M., Kumar, P.P., Han, J.H., Liou, Y.C., and Yu, H. (2008). Direct interaction of AGL24 and SOC1 integrates flowering signals in *Arabidopsis*. Development **135**:1481–1491.

Liu, C., Xi, W.Y., Shen, L.S., Tan, C.P., and Yu, H. (2009). Regulation of floral patterning by flowering time genes. Dev. Cell **16**:711–722.

Liu, X., Kim, Y.J., Müller, R., Yumul, R.E., Liu, C., Pan, Y., Cao, X., Goodrich, J., and Chen, X. (2011). *AGAMOUS* terminates floral stem cell maintenance in *Arabidopsis* by direct repressing *WUSCHEL* through recruitment of polycomb group proteins. Plant Cell **23**:3654–3670.

Liu, C., Teo, Z.W., Bi, Y., Song, S., Xi, W., Yang, X., Yin, Z., and Yu, H. (2013). A conserved genetic pathway determines inflorescence architecture in *Arabidopsis* and rice. Dev. Cell **24**:612–622.

Lohmann, J.U., Hong, R.L., Hobe, M., Busch, M.A., Parcy, F., Simon, R., and Weigel, D. (2001). A molecular link between stem cell regulation and floral patterning in *Arabidopsis*. Cell **105**:793–803.

Mandel, M.A., and Yanofsky, M.F. (1995). A gene triggering flower formation in *Arabidopsis*. Nature **377**:522–524.

Martínez-Castilla, L.P., and Álvarez-Buylla, E.R. (2003). Adaptive evolution in the *Arabidopsis* MADS-box gene family inferred from its complete resolved phylogeny. Proc. Natl. Acad. Sci. USA **100**: 13407–13412.

Masiero, S., Li, M.A., Will, I., Hartmann, U., Saedler, H., Huijser, P., Schwarz-Sommer, Z., and Sommer, H. (2004). INCOMPOSITA: a MADS-box gene controlling prophyll development and floral meristem identity in Antirrhinum. Development **131**:5981–5990.

Melzer, S., Lens, F., Gennen, J., Vanneste, S., Rohde, A., and Beeckman, T. (2008). Flowering-time genes modulate meristem determinacy and growth form in *Arabidopsis thaliana*. Nat. Genet. **40**:1489–1492.

Michaels, S.D., and Amasino, R.M. (1999). FLOWERING LOCUS C encodes a novel MADS domain protein that acts as a repressor of flowering. Plant Cell **11**:949–956.

Michaels, S.D., Ditta, G., Gustafson-Brown, C., Pelaz, S., Yanofsky, M., and Amasino, R.M. (2003). AGL24 acts as a promoter of flowering in *Arabidopsis* and is positively regulated by vernalization. Plant J. **33**:867–874.

Mizukami, Y., and Ma, H. (1997). Determination of *Arabidopsis* floral meristem identity by AGAMOUS. Plant Cell **9**:393–408.

Moon, J., Suh, S.S., Lee, H., Choi, K.R., Hong, C.B., Paek, N.C., Kim, S.G., and Lee, I. (2003). The SOC1 MADS-box gene integrates vernalization and gibberellin signals for flowering in *Arabidopsis*. Plant J. **35**:613–623.

Müller-Xing, R., Clarenz, O., Pokorny, L., Goodrich, J., and Schubert, D. (2014). Polycomb-group proteins and flowering locus to maintain commitment to flowering in *Arabidopsis thaliana*. Plant Cell **26**:2457–2471.

Murashige, T., and Skoog, F. (1962). A revised medium for rapid growth and bio assays with tobacco tissue cultures. Physiol. Plant. **15**: 473–497.

Ng, M., and Yanofsky, M.F. (2001). Activation of the *Arabidopsis* B class homeotic genes by APETALA1. Plant Cell **13**:739–753.

Ohshima, S., Murata, M., Sakamoto, W., Ogura, Y., and Motoyoshi, F. (1997). Cloning and molecular analysis of the *Arabidopsis* gene terminal flower 1. Mol. Gen. Genet. **254**:186–194.

Okamuro, J.K., Denboer, B.G.W., and Jofuku, K.D. (1993). Regulation of *Arabidopsis* flower development. Plant Cell **5**:1183–1193.

Okamuro, J.K., denBoer, B.G.W., LotysPrass, C., Szeto, W., and Jofuku, K.D. (1996). Flowers into shoots: photo and hormonal control of a meristem identity switch in *Arabidopsis*. Proc. Natl. Acad. Sci. USA **93**:13831–13836.

Okamuro, J.K., Szeto, W., LotysPrass, C., and Jofuku, K.D. (1997). Photo and hormonal control of meristem identity in the *Arabidopsis* flower mutants apetala2 and apetala1. Plant Cell **9**:37–47.

Parcy, F., Bomblies, K., and Weigel, D. (2002). Interaction of LEAFY, AGAMOUS and TERMINAL FLOWER1 in maintaining floral meristem identity in *Arabidopsis*. Development **129**:2519–2527.

Parenicová, L., de Folter, S., Kieffer, M., Horner, D.S., Favalli, C., Busscher, J., Cook, H.E., Ingram, R.M., Kater, M.M., Davies, B., et al. (2003). Molecular and phylogenetic analyses of the complete MADS-box transcription factor family in *Arabidopsis*: new openings to the MADS world. Plant Cell **15**:1538–1551.

Porri, A., Torti, S., Romera-Branchat, M., and Coupland, G. (2012). Spatially distinct regulatory roles for gibberellins in the promotion of flowering of *Arabidopsis* under long photoperiods. Development **139**:2198–2209.

Posé, D., Yant, L., and Schmid, M. (2012). The end of innocence: flowering networks explode in complexity. Curr. Opin. Plant Biol. **15**:45–50.

Putterill, J., Robson, F., Lee, K., Simon, R., and Coupland, G. (1995). The constans gene of *Arabidopsis* promotes flowering and encodes a protein showing similarities to zinc-finger transcription factors. Cell **80**:847–857.

Ratcliffe, O.J., Amaya, I., Vincent, C.A., Rothstein, S., Carpenter, R., Coen, E.S., and Bradley, D.J. (1998). A common mechanism controls the life cycle and architecture of plants. Development **125**:1609–1615.

Ratcliffe, O.J., Bradley, D.J., and Coen, E.S. (1999). Separation of shoot and floral identity in *Arabidopsis*. Development **126**:1109–1120.

Rieu, I., Ruiz-Rivero, O., Fernandez-Garcia, N., Griffiths, J., Powers, S.J., Gong, F., Linhartova, T., Eriksson, S., Nilsson, O., Thomas, S.G., et al. (2008). The gibberellin biosynthetic genes AtGA20ox1 and AtGA20ox2 act, partially redundantly, to promote growth and development throughout the *Arabidopsis* life cycle. Plant J. **53**: 488–504.

Roeder, A.H., and Yanofsky, M.F. (2006). Fruit development in *Arabidopsis*. *Arabidopsis* Book **4**:e0075.

Rounsley, S.D., Ditta, G.S., and Yanofsky, M.F. (1995). Diverse roles for MADS box genes in *Arabidopsis* development. Plant Cell **7**:1259–1269.

Roux, F., Touzet, P., Cuguen, J., and Le Corre, V. (2006). How to be early flowering: an evolutionary perspective. Trends Plant Sci. **11**:375–381.

Schmid, M., Uhlenhaut, N.H., Godard, F., Demar, M., Bressan, R., Weigel, D., and Lohmann, J.U. (2003). Dissection of floral induction pathways using global expression analysis. Development **130**:6001–6012.

Schönrock, N., Bouveret, R., Leroy, O., Borghi, L., Kohler, C., Gruissem, W., and Hennig, L. (2006). Polycomb-group proteins repress the floral activator AGL19 in the FLC-independent vernalization pathway. Genes Dev. **20**:1667–1678.

Schultz, E.A., and Haughn, G.W. (1991). Leafy, a homeotic gene that regulates inflorescence development in *Arabidopsis*. Plant Cell **3**:771–781.

Schultz, E.A., and Haughn, G.W. (1993). Genetic-analysis of the floral initiation process (Flip) in *Arabidopsis*. Development **119**:745–765.

Shannon, S., and Meeks-Wagner, D.R. (1991). A mutation in the *Arabidopsis* Tfl1 gene affects inflorescence meristem development. Plant Cell **3**:877–892.

Shannon, S., and Meeks-Wagner, D.R. (1993). Genetic interactions that regulate inflorescence development in *Arabidopsis*. Plant Cell **5**:639–655.

Sheldon, C.C., Rouse, D.T., Finnegan, E.J., Peacock, W.J., and Dennis, E.S. (2000). The molecular basis of vernalization: the central role of FLOWERING LOCUS C (FLC). Proc. Natl. Acad. Sci. USA **97**:3753–3758.

Simpson, G.G. (2004). The autonomous pathway: epigenetic and post-transcriptional gene regulation in the control of *Arabidopsis* flowering time. Curr. Opin. Plant Biol. **7**:570–574.

Smaczniak, C., Immink, R.G., Angenent, G.C., and Kaufmann, K. (2012). Developmental and evolutionary diversity of plant MADS-domain factors: insights from recent studies. Development **139**: 3081–3098.

Smyth, D.R., Bowman, J.L., and Meyerowitz, E.M. (1990). Early flower development in *Arabidopsis*. Plant Cell **2**:755–767.

Srikanth, A., and Schmid, M. (2011). Regulation of flowering time: all roads lead to Rome. Cell. Mol. Life Sci. **68**:2013–2037.

Suárez-López, P., Wheatley, K., Robson, F., Onouchi, H., Valverde, F., and Coupland, G. (2001). CONSTANS mediates between the circadian clock and the control of flowering in *Arabidopsis*. Nature **410**:1116–1120.

Sugimoto, K., Gordon, S.P., and Meyerowitz, E.M. (2011). Regeneration in plants and animals: dedifferentiation, transdifferentiation, or just differentiation? Trends Cell Biol. **21**:212–218.

Sun, B., and Ito, T. (2010). Floral stem cells: from dynamic balance towards termination. Biochem. Soc. Trans. **38**:613–616.

Sun, B., Xu, Y.F., Ng, K.H., and Ito, T. (2009). A timing mechanism for stem cell maintenance and differentiation in the *Arabidopsis* floral meristem. Genes Dev. **23**:1791–1804.

Tapia-López, R., Garcia-Ponce, B., Dubrovsky, J.G., Garay-Arroyo, A., Perez-Ruiz, R.V., Kim, S.H., Acevedo, F., Pelaz, S., and Álvarez-Buylla, E.R. (2008). An AGAMOUS-related MADS-box gene, XAL1 (AGL12), regulates root meristem cell proliferation and flowering transition in *Arabidopsis*. Plant Physiol. **146**:1182–1192.

Tooke, F., Ordidge, M., Chiurugwi, T., and Battey, N. (2005). Mechanisms and function of flower and inflorescence reversion. J. Exp. Bot. **56**:2587–2599.

Trevaskis, B., Tadege, M., Hemming, M.N., Peacock, W.J., Dennis, E.S., and Sheldon, C. (2007). Short vegetative phase-like MADS-box genes inhibit floral meristem identity in barley. Plant Physiol. **143**:225–235.

Vandesompele, J., De Preter, K., Pattyn, F., Poppe, B., Van Roy, N., De Paepe, A., and Speleman, F. (2002). Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. Genome Biol. **3**, research0034.

van Dijk, A.D., Morabito, G., Fiers, M., van Ham, R.C., Angenent, G.C., and Immink, R.G. (2010). Sequence motifs in MADS transcription factors responsible for specificity and diversification of protein-protein interaction. PLoS Comput. Biol. **6**:e1001017.

van Mourik, S., van Dijk, A.D.J., de Gee, M., Immink, R.G.H., Kaufmann, K., Angenent, G.C., van Ham, R.C.H.J., and Molenaar, J. (2010). Continuous-time modeling of cell fate determination in *Arabidopsis* flowers. BMC Syst. Biol. **4**:101.

Villarreal, C., Padilla-Longoria, P., and Álvarez-Buylla, E.R. (2012). General theory of gene to phenotype mapping: derivation of epigenetic landscapes from n-node complex gene regulatory networks. Phys. Rev. Lett. **109**:118102.

Wang, J.W., Czech, B., and Weigel, D. (2009). miR156-regulated SPL transcription factors define an endogenous flowering pathway in *Arabidopsis thaliana*. Cell **138**:738–749.

Weigel, D., Alvarez, J., Smyth, D.R., Yanofsky, M.F., and Meyerowitz, E.M. (1992). Leafy controls floral meristem identity in *Arabidopsis*. Cell **69**:843–859.

Wigge, P.A., Kim, M.C., Jaeger, K.E., Busch, W., Schmid, M., Lohmann, J.U., and Weigel, D. (2005). Integration of spatial and temporal information during floral induction in *Arabidopsis*. Science **309**:1056–1059.

Wilkinson, D.J. (2011). Stochastic Modelling for Systems Biology, 2nd edn (Boca Raton, FL, USA: Chapman & Hall/CRC Mathematical and Computational Biology), p. 363.

Wu, G., and Poethig, R.S. (2006). Temporal regulation of shoot development in *Arabidopsis thaliana* by miR156 and its target SPL3. Development **133**:3539–3547.

Yamaguchi, A., Wu, M.F., Yang, L., Wu, G., Poethig, R.S., and Wagner, D. (2009). The microRNA-regulated SBP-box transcription factor SPL3 is a direct upstream activator of LEAFY, FRUITFULL, and APETALA1. Dev. Cell **17**:268–278.

Yu, H., Xu, Y.F., Tan, E.L., and Kumar, P.P. (2002). AGAMOUS-LIKE 24, a dosage-dependent mediator of the flowering signals. Proc. Natl. Acad. Sci. USA **99**:16336–16341.

Yu, H., Ito, T., Wellmer, F., and Meyerowitz, E.M. (2004). Repression of AGAMOUS-LIKE 24 is a crucial step in promoting flower development. Nat. Genet. **36**:157–161.

Zhao, S., and Fernald, R.D. (2005). Comprehensive algorithm for quantitative real-time polymerase chain reaction. J. Comput. Biol. **12**:1047–1064.

Zhou, J.X., Qiu, X., Fouquier d'Hérouël, A., and Huang, S. (2014). Discrete gene network models for understanding multicellularity and cell reprogramming: from network structure to attractor landscape. In Computational Systems Biology, 2nd edn, R. Eils and A. Kriete, eds. (San diego, CA, USA: Elsevier), pp. 241–276.

Chapter 5

Conclusiones

In spite of its familiarity, the formation of plausible conclusions is a very subtle process.
— E . T . Jaynes, *Probability Theory - The Logic of Science* (2003)

En este proyecto se presenta la perspectiva de un modelo de mapeo genotipo a fenotipo en términos del rol auto-organizacional de redes regulatorias genéticas para abordar el problema general de la decisión del destino celular. Se argumenta con base en esta perspectiva que modelos extendidos de redes regulatorias genéticas pueden representar efectivamente un Paisaje Epigenético subyacente a un proceso de desarrollo. La caracterización de las propiedades estructurales y cuantitativas de este paisaje pueden ayudar tanto a entender como a predecir eventos celulares durante procesos de desarrollo, y potencialmente la evolución de estos últimos.

De manera concreta, se propone un marco metodológico para extender modelos de redes regulatorias genéticas con la intensión de investigar el impacto de perturbaciones a genes específicos en la toma de decisión celular como resultado de la re–estructuración del Paisaje Epigenético subyacente (*Articulo VI*). Mediante la aplicación del marco metodológico al caso práctico del desarrollo floral, se muestra que el Paisaje Epigenético puede ser re-estructurado mediante la modulación de los tiempos característicos de expresión de genes particulares, y se sugiere que este fenómeno es importante para entender de manera mecanicista el funcionamiento interno de las células durante la decisión sobre su destino. Los resultados obtenidos sugieren que existe una relación entre el impacto de genes específicos en la dinámica de la red regulatoria genética, su rol biológico y la observación jerárquica de eventos de decisión celular durante el desarrollo temprano de la flor. Adicionalmente se especula que el rol dinámico diferencial de los genes

descubierto aquí podría dar información sobre la tendencia de los genes para ligar el módulo regulatorio con otros circuitos regulatorios o vías de transducción de seales.

En un segundo modelo se integraron datos experimentales en un modelo integrativo de red de regulación genética. Se propone que la red obtenida constituye un modelo genérico para el proceso de transformación tumorigénica potencial observado in-vitro y descrito somo el proceso de inmortalización espontanea (*Artículo VII*). Mediante el análisis dinámico de la red y su Paisaje Epigenético subyacente se presenta evidencia de que los componentes moleculares y las interacciones consideradas son necesarios y suficientes para recuperar los destinos celulares y transiciones observadas durante el fenómeno biológico. Cabe destacar que los destinos celulares recuperados con en el modelo, y su patrón de transiciones, correlaciona con los patrones observados durante la progresión de la carcinogenesis epitelial in vivo, esto evidenciado por descripciones patológicas. Los resultados presentados sugieren, entonces, que la potencial transformación tumorigénica in-vitro como resultado del proceso de inmortalización espontanea es adecuadamente entendido y modelado al nivel celular de manera genérica como un sistema en desarrollo que presenta decisiones del destino celular como resultado de las restricciones estructurales y funcionales impuestas, en parte, por las interacciones incluidas en la red subyacente propuesta.

Por último, bajo la hipótesis de que la relevancia funcional de un red regulatoria subyacente a un proceso de desarrollo impide una alto grado de variación durante la evolución, en este proyecto se prueba que los componentes de tal red involucrada en el establecimiento de los destinos celulares durante el desarrollo temprano de la flor de Arabidopsis se encuentran conservados a nivel molecular a lo largo de 18 especies de plantas con flor (*Articulo IX*). Adicionalmente, se prueba que existe evidencia de que la red regulatoria ha sido sometida a restricciones funcionales durante la evolución. Los resultados presentados aquí soportan la hipótesis original de que la red regulatoria estudiada constituye un módulo regulatorio que regula un proceso de desarrollo de manera robusta y que ha sido sometido a fuertes restricciones funcionales durante la evolución.

En el proyecto en su totalidad presentamos antecedentes necesarios y propuestas de modelado específicas para sustanciar nuestra conclusión de que el conjunto de modelos definidos aquí como el Paisaje Epigenético de Atractores (*Artículo* V), se perfilan como la extensión más

natural para continuar el protocolo básico de modelado de redes regulatorias genéticas y asi extender el enfoque de biología de sistemas en el estudio del desarrollo. Por último, para impulsar esta adición al modelado en biología de sistemas, se presenta aquí una implementación novedosa de los métodos de modelaje del Paisaje Epigenético de Atractores asociado a redes regulatorias genéticas que esperamos será de utilidad para la comunicad científica en la interface entre la biología y las disciplinas cuantitativas (*Articulo VIII*).

Bibliography

Azpeitia, E., Davila-Velderrain, J., Villarreal, C., y Alvarez-Buylla, E.R. Gene regulatory network models for floral organ determination. En *Flower Development*, págs. 441–469. Springer (2014)

Davila-Velderrain, J., Martinez-Garcia, J., y Alvarez-Buylla, E. Descriptive vs. Mechanistic Network Models in Plant Development in the Post-Genomic Era. *Plant Functional Genomics: Methods and Protocols* págs. 455–479 (2015a)

Dávila-Velderrain, J. y Álvarez-Buylla Roces, E. Linear Causation Schemes in Postgenomic Biology: The Subliminal and Convenient One-to-one Genotype-Phenotype Mapping Assumption. *INTERdisciplina* **3**(5) (????)

Davila-Velderrain, J., Martinez-Garcia, J.C., y Alvarez-Buylla, E.R. Epigenetic Landscape Models: The Post-Genomic Era. *bioRxiv* (2014a)

Davila-Velderrain, J., Servin-Marquez, A., y Alvarez-Buylla, E.R. Molecular evolution constraints in the floral organ specification gene regulatory network module across 18 angiosperm genomes. *Molecular biology and evolution* **31**(3):560–573 (2014b)

Davila-Velderrain, J., Martínez-García, J., y Alvarez-Buylla, E.R. Modeling the Epigenetic Attractors Landscape: Towards a Post-Genomic Mechanistic Understanding of Development. *Name: Frontiers in Genetics* **6**:160 (2015b)

Pérez-Ruiz, R.V., García-Ponce, B., Marsch-Martínez, N., Ugartechea-Chirino, Y., Villajuana-Bonequi, M., de Folter, S., Azpeitia, E., Dávila-Velderrain, J., Cruz-Sánchez, D., Garay-Arroyo, A. *et al.* XAANTAL2 (AGL14) is an important

component of the complex gene regulatory network that underlies arabidopsis shoot apical meristem transitions. *Molecular plant* **8**(5):796–813 (2015)