



UNIVERSIDAD NACIONAL AUTÓNOMA DE  
MÉXICO

DOCTORADO EN CIENCIAS BIOMÉDICAS  
INSTITUTO DE ECOLOGÍA

Investigación de propiedades estructurales y dinámicas subyacentes a  
la troncalidad: Una perspectiva de sistemas complejos a la biología  
de células troncales embrionarias

T E S I S

QUE PARA OPTAR POR EL GRADO DE  
DOCTOR EN CIENCIAS

PRESENTA:  
JOSÉ LUIS CALDÚ PRIMO

*DIRECTORA DE TESIS*  
DRA. MARÍA ELENA ÁLVAREZ-BUYLLA ROCES  
INSTITUTO DE ECOLOGÍA

*COMITÉ TUTOR*  
DR. CARLOS VILLARREAL LUJÁN  
INSTITUTO DE FÍSICA  
DR. JUAN CARLOS MARTÍNEZ GARCÍA  
DEPARTAMENTO DE CONTROL AUTOMÁTICO, CINVESTAV

CIUDAD UNIVERSITARIA, CD.MX., México OCTUBRE 2021



Universidad Nacional  
Autónoma de México



**UNAM – Dirección General de Bibliotecas**  
**Tesis Digitales**  
**Restricciones de uso**

**DERECHOS RESERVADOS ©**  
**PROHIBIDA SU REPRODUCCIÓN TOTAL O PARCIAL**

Todo el material contenido en esta tesis esta protegido por la Ley Federal del Derecho de Autor (LFDA) de los Estados Unidos Mexicanos (México).

El uso de imágenes, fragmentos de videos, y demás material que sea objeto de protección de los derechos de autor, será exclusivamente para fines educativos e informativos y deberá citar la fuente donde la obtuvo mencionando el autor o autores. Cualquier uso distinto como el lucro, reproducción, edición o modificación, será perseguido y sancionado por el respectivo titular de los Derechos de Autor.



# Agradecimientos

A la UNAM, en donde me he formado durante los últimos 11 años.

A José, con quien hice casi todo el trabajo de esta tesis por su guía y amistad.

A Elena, por confiar en mí, apoyarme y guiarme.

A Juan Carlos y a Carlos por sus consejos, las pláticas y su amistad.

A mis sinodales, Rosana Pelayo, Julián Valdés, Luis Mendoza y Alfredo Varela, por su revisión de esta tesis.

Al CONACyT, por darme una beca que me permitió estudiar este posgrado.

A Antonio Scialdone y los demás miembros del Instituto Helmholtz de Munich, que me recibieron durante un semestre.

A mis papás.

A Anahí, Tania, Thomas, Churro, Maia, Mati y Ánika, que siempre están conmigo a pesar de la distancia.

A Isa y Paco.

A mis compañeros del laboratorio y del C3, en particular, a Moni, Caro, Jenny, Stalin, Mariana Esther, Elisa, Juan y al Doc. Verduzco.

A mis tíos y primos.

A mis amigos de la Facultad, del Moderno, del Jardín y los demás.

A Fersi.



# Resumen

La troncalidad, entendida como la capacidad celular de auto-renovación y diferenciación en tipos celulares más especializados, es un concepto central en biología del desarrollo. Un mejor entendimiento de los fundamentos de la troncalidad es esencial para comprender el desarrollo de organismos multicelulares y el surgimiento de enfermedades degenerativas. Por ello, distintas ramas de la biología han buscado explicaciones de la troncalidad, desde caracterizaciones moleculares a planteamientos teóricos. Las células troncales se caracterizan por tener un alto grado de heterogeneidad transcripcional; lo que dificulta una definición basada en perfiles estables de expresión genética. Desde una perspectiva epigenómica, las células troncales se distinguen por tener un perfil de cromatina ‘abierta’, con una reducción de marcas represivas de histonas en comparación con células diferenciadas. Asimismo, planteamientos teóricos basados en teoría de sistemas dinámicos, han propuesto que las propiedades de auto-renovación y diferenciación implican un balance entre robustez y plasticidad en el sistema regulatorio subyacente. Sin embargo, a pesar la acumulación de conocimiento sobre la troncalidad desde diferentes perspectivas, la integración de ideas teóricas con evidencias experimentales bajo un marco conceptual común es un reto pendiente todavía. Motivado por esta cuestión, he hecho contribuciones al entendimiento teórico de la troncalidad por medio de modelos dinámicos de redes de regulación genética, análisis estructurales de redes de regulación específicas de tipo celular, evaluación funcional de genes esenciales y una propuesta para la interpretación teórica de la información epigenómica. Esta tesis es la integración de estos trabajos.

Los modelos de redes dinámicas de regulación genética son un método consolidado para el estudio de la diferenciación celular. Debido a esto, comienzo con una revisión técnica de este tipo de modelos y examino la conveniencia de usarlos para el estudio del cáncer. Continúo con una caracterización estructural de redes de factores de transcripción específicas de tipo celular, inferidas a partir de datos empíricos de accesibilidad a la cromatina. Este análisis revela propiedades fundamentales distintivas de las células troncales, que plantean un fundamento sistémico para atributos biológicos asociados a éstas. A continuación, presento un estudio bioinformático de genes esenciales, definidos como tal por ser intolerantes a las mutaciones en ámbitos *in vitro* o *in vivo*. Este análisis demuestra la dependencia de contexto en la función de los genes y en particular la necesidad de estudiar genes relacionados con el desarrollo tomando en cuenta su función a nivel organismo. Finalmente, presento ideas para la integración de la información epigenómica dentro de la teoría del paisaje epigenético, que representa la diferenciación celular como un sistema dinámico complejo. Considerando a las modificaciones epigenéticas parte del sistema complejo subyacente a la diferenciación celular, los cambios epigenómicos que suceden durante el proceso de desarrollo pueden ser interpretados en términos de las consecuencias dinámicas que tienen en la red regulatoria subyacente.

Por medio de la integración de perspectivas empíricas y acercamientos teóricos del estudio del desarrollo, esta tesis busca ser una aportación hacia el mejor entendimiento de la troncalidad.



# Abstract

Stemness, understood as a cell's capacity to self-renew and differentiate into more specialized cell types, is a central concept in developmental biology. Developing a deeper understanding of the biological basis of stemness is essential to better comprehend the development of multicellular organisms and the onset of degenerative diseases. Therefore, several branches of biology have proposed explanations for stemness, from molecular characterizations to theoretical deliberations. Stem cells are characterized by having a high degree of transcriptional heterogeneity; which prohibits simple definitions based on stable gene expression profiles. From an epigenomic perspective, stem cells are distinguished by an 'open' chromatin profile, with a depletion in repressive histone marks compared to differentiated cells. Furthermore, theoretical approaches based on dynamical systems theory, have proposed that the properties of self-renewal and differentiation imply a balance between robustness and plasticity in the underlying regulatory system. Despite this accumulation of knowledge on stemness from different perspectives, however, the integration of theoretical ideas with experimental evidence under a common conceptual framework is still a long-standing problem. Motivated by this challenge, I have made contributions to the theoretical understanding of stemness by means of dynamical models of gene regulatory networks, structural analysis of empirical cell type specific regulatory networks, functional examination of gene essentiality, and a proposal for the theoretical interpretation of epigenomic information. This thesis is the integration of this work.

Dynamical gene regulatory network models are a widespread method to study cell differentiation. Therefore, I start with a technical review of this kind of models, and examine the convenience of using them for the study of cancer. Having set the theoretical basis for cell differentiation, I continue with a structural characterization of cell type specific transcription factor networks inferred from empirical chromatin accessibility data. This analysis uncovers fundamental properties distinctive of stem cells, that offer a systemic foundation for biological features associated with stemness. Next, I present a bioinformatic analysis of essential genes, defined as such for their intolerance to mutations at *in vivo* or *in vitro* settings. This study demonstrates the context dependence of gene function, and particularly the necessity of taking into account the organismal level when studying genes with a developmental function. Finally, I present ideas for the integration of epigenomic information into the epigenetic landscape theory, that represents cell differentiation as a complex dynamical system. Considering epigenetic modifications a part of the complex dynamical system underlying cell differentiation, the epigenomic profile can be interpreted in terms of the dynamical consequences it has on the underlying regulatory network.

Through the integration of empirical and theoretical perspectives to the study of development, this thesis aims to contribute to a better understanding of stemness.





# Índice general

<b>Agradecimientos</b>	<b>I</b>
<b>Resumen</b>	<b>III</b>
<b>Abstract</b>	<b>V</b>
<b>Esquema general de la tesis</b>	<b>IX</b>
<b>1. Introducción</b>	<b>1</b>
1.1. Las células troncales, modelo del desarrollo celular . . . . .	1
1.2. La troncalidad vista desde el paisaje epigenético . . . . .	5
1.3. Una aproximación integrativa a la troncalidad . . . . .	7
<b>2. Planteamiento del problema y objetivos</b>	<b>9</b>
<b>3. Modelado del paisaje epigenético de atractores</b>	<b>11</b>
3.1. Modelado computacional del paisaje epigenético . . . . .	12
3.2. Estudiando el cáncer a través del paisaje epigenético . . . . .	15
3.3. La troncalidad y el cáncer . . . . .	16
<b>4. Propiedades estructurales de la red de factores de transcripción de células troncales embrionarias</b>	<b>19</b>
<b>5. Análisis de la dependencia del contexto en la definición de genes ‘esenciales’ en humano</b>	<b>23</b>
<b>6. El perfil epigenómico visto desde el paisaje epigenético de Waddington</b>	<b>27</b>
<b>7. Discusión y conclusiones</b>	<b>31</b>
<b>8. Perspectivas a futuro</b>	<b>35</b>
<b>Apéndices</b>	<b>37</b>
<b>A. Modeling the epigenetic landscape in plant development</b>	<b>39</b>
<b>B. Cancer in attractor landscape modeling: a systems biology perspective of the disease</b>	<b>67</b>

- C. Structural robustness of mammalian transcription factor networks reveals plasticity across development 83
- D. In vivo and in vitro human gene essentiality estimations capture constrasting functional constraints 99
- E. Understanding epigenomics in the context of Waddington's epigenetic landscape 115

# Esquema general de la tesis

- **1. Introducción**
- **2. Planteamiento del problema y objetivos**
- **3. Modelado del paisaje epigenético de atractores**

Este capítulo presenta una revisión del modelado computacional del paisaje epigenético de atractores que es usado como base teórica para el estudio del desarrollo. Este capítulo está basado en dos artículos en los que participé como co-autor y que han sido publicados como capítulos en libros de divulgación científica. Los artículos son:

1. Davila-Velderrain, J., Caldu-Primo, J. L., Martinez-Garcia, J. C., and Alvarez-Buylla, E. R. (2018) *Modeling the epigenetic landscape in plant development*. In von Stechow, L. and Santos Delgado, A., (eds.), *Computational Cell Biology*, Vol. 1819, pp. 357–383 Springer International Publishing.
2. Caldu-Primo, J. L., Davila-Velderrain, J., Martinez-Garcia, J. C., and Alvarez-Buylla, E. R. (2018) *Cancer in attractor landscape modeling: A systems biology perspective of the disease*. In Miramontes Vidal, O. R. and Alvarez-Buylla, E. R., (eds.), *Cancer a complex disease*, pp. 3–18 CopIt-arXives.

El primer artículo presenta una revisión técnica del modelado estocástico del paisaje de atractores a partir de redes Booleanas de regulación genética, usando como ejemplo las redes de diferenciación celular en raíz de *Arabidopsis thaliana*. El segundo artículo es una revisión de modelos de redes dinámicas para estudiar la aparición de cáncer en diferentes tejidos. Esta revisión plantea algunas ventajas que se obtienen al abordar una enfermedad asociada al desarrollo, como el cáncer, desde la perspectiva sistémica del paisaje epigenético.

- **4. Propiedades estructurales de la red de células troncales embrionarias**

Este capítulo presenta los resultados del análisis estructural de redes de factores de transcripción específicas de tipo celular que fueron publicados en la revista *Scientific Reports* en el artículo:

3. Caldu-Primo, J. L., Alvarez-Buylla, E. R., and Davila-Velderrain, J. (dec,2018) *Structural robustness of mammalian transcription factor net-*

*works reveals plasticity across development*. Scientific Reports, 8(1), 1–15.

La base del artículo es una comparación estructural de redes de regulación, construidas a partir de datos epigenómicos, que muestra una arquitectura contrastante entre células troncales embrionarias y tipos celulares diferenciados. El análisis de redes permite hacer una interpretación de los datos epigenómicos desde la óptica de sistemas dinámicos y vincularlos al marco teórico del paisaje epigenético.

#### ■ 5. Dependencia de contexto en la definición de genes esenciales

Este capítulo presenta un análisis de propiedades asociadas a genes intolerantes a las mutaciones, comparándolos de acuerdo al contexto en el que fueron definidos: *in vitro* o *in vivo*. Este análisis forma parte de un artículo publicado en *Nucleic Acids Research: Genomics and Bioinformatics*:

4. Caldu-Primo, J. L., Verduzco-Martínez, J. A., Alvarez-Buylla, E. R., and Davila-Velderrain, J. (2021) *In vivo and in vitro human gene essentiality estimations capture contrasting functional constraints*. NAR Genomics and Bioinformatics, 3(3), 1–14. <https://doi.org/10.1093/nargab/lqab063>

Este estudio resalta inconvenientes al usar la perspectiva genocéntrica en biología y la necesidad de considerar el contexto en el que se presenta un gen para definir su función. Desde este punto de vista, tiene relevancia para el estudio del desarrollo y de la troncalidad al resaltar que estudios *in vitro* no siempre son útiles para analizar genes involucrados en el desarrollo ya que su función depende de las interacciones que mantengan con el resto del sistema.

#### ■ 6. El perfil epigenómico interpretado desde el paisaje epigenético de Waddington

En este último capítulo presento una propuesta de la forma en que se podría incorporar la información epigenómica al marco teórico del paisaje epigenético de atractores. Estas ideas forman parte de un ensayo en proceso de publicación:

5. Epigenomics in the context of Waddington's epigenetic landscape

#### ■ 7. Discusión y conclusiones

#### ■ 8. Perspectivas a futuro

# 1 Introducción

## 1.1. Las células troncales, modelo del desarrollo celular

Una célula troncal es una célula capaz de auto-renovarse, dando lugar a más células troncales, y de diferenciarse en tipos celulares más especializados (Figura 1.1) [1–3]. Siguiendo esta definición, dentro del concepto de troncalidad esta contenida una pregunta básica de la biología del desarrollo: ¿cómo puede una célula dar lugar a cientos de tipos celulares diferentes a partir de la misma información genética? [4] Como tal, la troncalidad es un concepto íntimamente asociado al desarrollo y a la diferenciación celular, por lo que para su comprensión se requiere del estudio de los procesos que median entre genotipo y fenotipo.

La definición de célula troncal a partir de dos características funcionales parece sencilla, sin embargo para entender realmente qué es una célula troncal se necesitan plantear algunas puntualizaciones. Antes que nada, vale la pena considerar qué es un tipo celular y si las células troncales pueden ser consideradas como tal. A mediados del siglo XIX Schleiden y Schwann plantearon la teoría celular, estableciendo que la célula es la unidad estructural y funcional básica de la vida, que todos los organismos están compuestos de células y que toda célula proviene de células preexistentes [5, 6]. A partir de entonces, ha existido la ambición de caracterizar y clasificar a las células que conforman a un organismo en distintos tipos celulares, basándose para esto en descripciones cada vez más detalladas de las células. La descripción y definición de tipos celulares ha sido guiada por el desarrollo tecnológico, teniendo avances críticos con el desarrollo de técnicas de microscopía, tinción celular, marcaje inmunológico y, durante las últimas décadas, técnicas moleculares de caracterización transcripcional [7]. Si bien no existe una definición rigurosa de ‘tipo celular’, de manera operacional se puede definir como un conjunto de células con un fenotipo estable que las distingue de otros tipos celulares. Este fenotipo está asociado a un perfil estable de expresión genética, en particular a la expresión de un conjunto de genes distintivos conocidos como genes marcadores [7–9].

Las células troncales, en general, no pueden ser consideradas un tipo celular debido a que en realidad el término engloba a un conjunto de tipos celulares diferentes que comparten las capacidades de auto-renovación y de diferenciación en células más especializadas, a esta última propiedad se le conoce como potencial de desarrollo [2, 3]. Los diferentes tipos de células troncales que existen se clasifican de acuerdo a su potencial de desarrollo, determinado por la cantidad de tipos celulares diferentes que pueden generarse a partir de ellas, en: células troncales totipotentes, pluripotentes, multipotentes y unipotentes. Las células troncales to-

tipotentes pueden diferenciarse en todos los tejidos extraembrionarios y todos los tejidos del cuerpo. Las células troncales pluripotentes son capaces de diferenciarse en cualquier célula de los tres tejidos embrionarios (endodermo, mesodermo y ectodermo). En cambio, las células multipotentes son capaces de diferenciarse únicamente en un subconjunto de tipos celulares del tejido en el que se encuentran y las células unipotentes sólo se diferencian en un tipo celular especializado. Además de su potencial de desarrollo, las células troncales se clasifican de acuerdo al tejido del organismo en el que se encuentran o la forma en que se obtienen. De este modo, el ejemplo clásico de una célula totipotente es el cigoto. Las células pluripotentes pueden ser células troncales embrionarias, obtenidas de la masa interna del blastocisto, o células troncales pluripotentes inducidas, obtenidas a partir de la reprogramación de una célula diferenciada. En cuanto a las células multipotentes y unipotentes, se han descrito una variedad de tipos celulares que residen en tejidos del organismo adulto y contribuyen a su funcionamiento y mantenimiento, algunos ejemplos son células troncales hematopoyéticas, progenitores neuronales, células troncales epidérmicas y espermatogonias [3, 10, 11]. Dentro de la variedad de células troncales que existen, las células pluripotentes son de especial interés en biología del desarrollo y medicina regenerativa debido a su capacidad de generar cualquier tipo celular del organismo adulto [2, 12]. Durante el resto de esta tesis, me enfocaré en las células troncales pluripotentes, siendo éstas el modelo ideal para abordar el estudio de la troncalidad.

La búsqueda de una explicación para el fenómeno de la troncalidad unida a la perspectiva genocéntrica en biología han promovido el interés en encontrar una base genética de la troncalidad independiente del linaje celular. Con este fin, se han realizado investigaciones buscando los “genes de la troncalidad”, refiriéndose a genes con expresión compartida entre diferentes células troncales y que las distinguen de células diferenciadas a partir de ellas [2, 13]. A partir esta idea, tres grupos de investigación independientes analizaron genes con expresión diferencial entre tres tipos de células troncales (células troncales embrionarias, progenitores neurales, células troncales hematopoyéticas y células troncales retinales) y células diferenciadas a partir de ellas [14–16]. Comparando las listas de genes que distinguían a las células troncales, cada estudio obtuvo una lista de alrededor de 300 genes en común entre los distintos tipos de células troncales. Sin embargo, al comparar las listas obtenidas en los diferentes estudios, solamente había un gen común (*Integrin Alfa 6*) que codifica a un receptor de membrana, lo cuál no se relaciona directamente con las funciones asociadas a la troncalidad, además de que esta intersección no es estadísticamente significativa. Este resultado debilitó la idea de que la troncalidad estuviera sustentada por la expresión de un grupo de genes [8, 16, 17]. Aunque no existan genes que expliquen a la troncalidad en general, se han encontrado genes marcadores para células troncales en linajes celulares particulares. Estos genes sirven para identificar a un tipo de célula troncal, sin embargo al no ser comunes a todas las células troncales no pueden considerarse una explicación genérica de la troncalidad. Un ejemplo de esto es la expresión de Oct4, Sox2 y Nanog en células troncales embrionarias [18].

Una consecuencia del fracaso en la búsqueda de una base genética para la troncalidad, ha sido considerarla un estado celular, definido como un atributo transitorio que puede adquirir cualquier célula y que no puede ser descrito por patrones de expresión celular [13]. Una ventaja del concepto de estado sobre el de

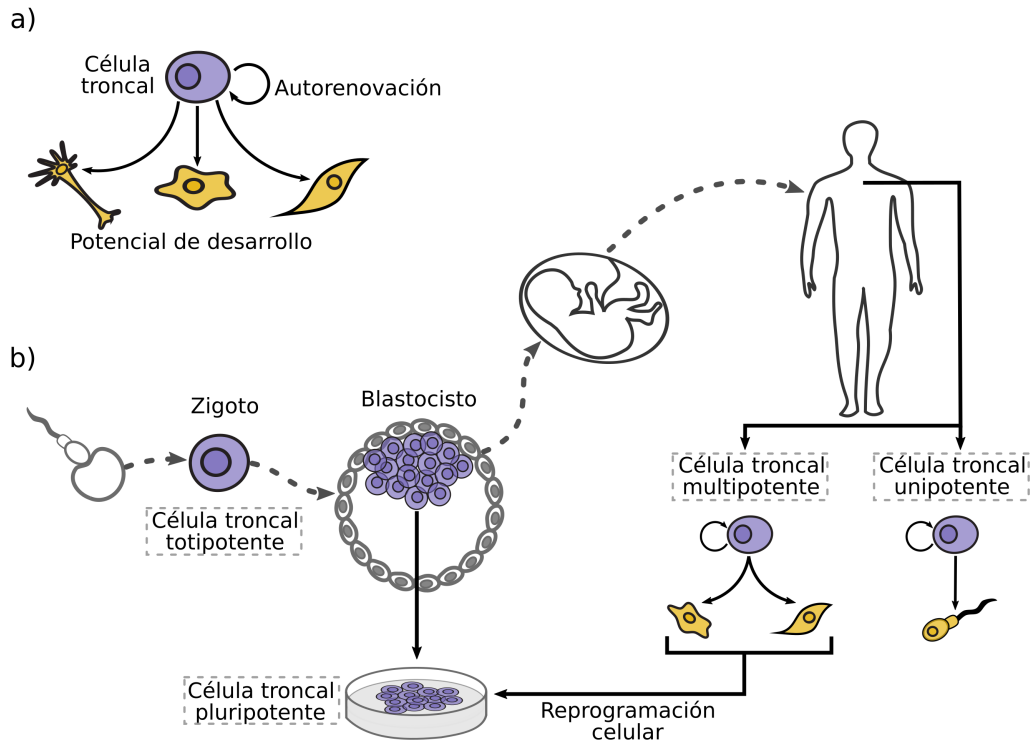


Figura 1.1: a) Una célula troncal está definida por su capacidad de auto-renovación y por su potencial de desarrollo para diferenciarse en tipos celulares más especializados. b) Las células troncales se clasifican por su potencial de desarrollo en totipotentes, pluripotentes, multipotentes y unipotentes. Estos tipos de células troncales se encuentran en diferentes etapas y partes del cuerpo humano.

tipo celular es que implica la idea de transitoriedad y reversibilidad de la troncalidad, y más en general del desarrollo [2, 19]. Sin embargo, en términos prácticos considerar a la troncalidad como un estado celular ofrece pocas ventajas en cuanto a la identificación de células troncales, para lo cual se sigue dependiendo de la caracterización funcional de éstas en cuanto a su capacidad de auto-renovación y potencial de desarrollo. Para esto se realizan experimentos en los que se extrae una muestra de tejido donde se encuentran las supuestas células troncales, esta muestra es colocada en un medio de cultivo que inhibe la diferenciación para probar la capacidad de auto-regeneración y después se vuelve a manipular experimentalmente para comprobar su potencial de desarrollo [2, 3, 20]. Existen diferentes experimentos para determinar el potencial de desarrollo de una célula, algunos de estos son: diferenciación *in vitro*, formación de teratomas y formación de quimeras al introducirlas a un embrión [3, 20].

El proceso de determinación funcional de la troncalidad hace imposible determinar ambas características para una misma célula, ya que al probar la auto-regeneración la célula se divide y por lo tanto la diferenciación se realiza en una célula distinta a la original. Esta cuestión ha sido nombrada por algunos autores como el principio de indeterminación [2]. Una forma de superar este problema sería considerar que todas las células de una población de células troncales son idénticas y al dividirse generan células exactamente iguales a la original. Sin embargo, las



poblaciones de células troncales se caracterizan por tener un alto grado de heterogeneidad, lo que impide el argumento de identidad entre las células [8, 13, 21, 22].

La heterogeneidad transcripcional se refiere a que las células que conforman a una población troncal tienen una alta variación en los niveles de expresión genética. La relación entre la heterogeneidad transcripcional y la troncalidad ha sido ampliamente estudiado desde que se observaron fluctuaciones en los niveles de expresión de *Nanog* y otros factores de pluripotencia en células troncales embrionarias de ratón [8, 23]. Más recientemente, este fenómeno ha sido estudiado mediante el uso de técnicas de caracterización transcripcional de célula única (scRNA-seq), confirmando la alta heterogeneidad transcripcional de células troncales [24, 25]. Debido a esto, la heterogeneidad transcripcional ha sido asociada al estado de indeterminación de las células troncales [22, 26]. Según esta idea, la heterogeneidad es una consecuencia del potencial de desarrollo, el cual implica una ausencia de compromiso hacia cualquiera de los linajes en los que se puede diferenciar una célula troncal y consecuentemente la expresión de genes asociados a los diferentes linajes. La ausencia de compromiso celular, a su vez, está asociada a un estado de regulación laxo que permite transitar entre diferentes perfiles de expresión, generando fluctuaciones en los niveles de expresión [1, 27, 28]. Bajo esta idea, la troncalidad, más que ser un atributo de una célula, es un fenotipo robusto a nivel poblacional que emerge a partir de dinámicas estocásticas al nivel de células únicas. Para entender la manera en que comportamientos robustos a nivel poblacional resultan de dinámicas estocásticas de los elementos subyacentes, se ha estudiado la troncalidad utilizando herramientas de mecánica estadística [8, 27, 29, 30].

El interés por entender a las células troncales, asociado al desarrollo de tecnologías que otorgan una capacidad de descripción molecular cada vez más detallada de una célula, ha llevado a la caracterización de células troncales pluripotentes en cuanto a su perfil epigenómico. Las modificaciones epigenéticas (metilación de DNA, modificaciones en histonas, arquitectura de la cromatina, entre otros) son mecanismos de control transcripcional que, en parte, determinan la expresión genética [31, 32]. Como tal, han sido estudiadas en células pluripotentes como posibles explicaciones moleculares para la troncalidad. A nivel de cromatina, las células troncales embrionarias comparadas con células diferenciadas tienen un perfil epigenómico con menos marcas represivas y más marcas de activación transcripcional [33, 34]. En particular, las células troncales embrionarias tienen relativamente poca metilación del DNA, relativamente pocas modificaciones en histonas y un mayor accesibilidad de la cromatina [33–37]. Una característica particular de las células pluripotentes es la presencia de modificaciones de histonas bivalentes, esto quiere decir que en una misma histona se encuentran marcas de activación y de represión de la transcripción [38, 39]. Estas características epigenómicas han sido asociadas al potencial de desarrollo de las células troncales y como tal a su estado de indeterminación y a la heterogeneidad transcripcional [37, 40].

La capacidad tecnológica de caracterización molecular ha llevado a una descripción cada vez más detallada de las células troncales y a un mayor conocimiento de los mecanismos moleculares que les permiten mantenerse en un estado indiferenciado. Sin embargo, para alcanzar una comprensión del fenómeno de la troncalidad, estos conocimientos tienen que estar sustentados en una teoría del desarrollo biológico que dé sentido a las diferentes características asociadas a la troncalidad [41]. Durante las últimas décadas, el desarrollo biológico ha sido exitosamente aborda-

do desde una perspectiva sistémica presentada primero metafóricamente y después formalmente mediante la imagen del paisaje epigenético (Figura 1.2).

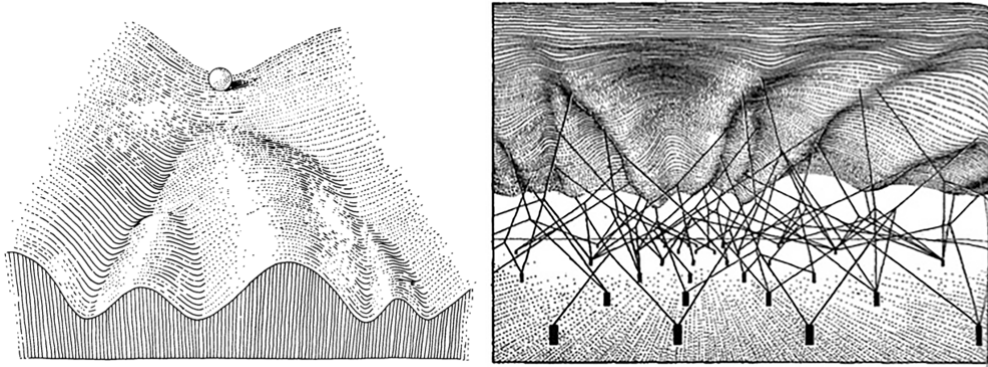


Figura 1.2: El paisaje epigenético propuesto por Waddington como metáfora del desarrollo muestra una esfera descendiendo por terreno en el que los valles se bifurcan. La esfera representa a células de un organismo que conforme avanza el desarrollo se diferencian tomando caminos diferentes y llegando a destinos celulares distintos. En la imagen de la derecha se muestra la idea de una red de genes subyacente que da forma al paisaje epigenético. Imágenes tomadas de [42].

## 1.2. La troncalidad vista desde el paisaje epigenético

Como se mencionó antes, uno de los objetivos principales de la biología del desarrollo es entender cómo pueden generarse cientos de fenotipos celulares diferentes a partir de la misma información genética. La existencia de diferentes tipos celulares en un organismo, compartiendo genoma pero con morfología y comportamiento completamente distintos, hace insostenible el mapeo uno a uno entre genotipo y fenotipo. Esto obliga a dejar de lado explicaciones reduccionistas del comportamiento celular basadas en cadenas lineales de causalidad genética, para pasar a una perspectiva sistémica que tome en cuenta las interacciones no lineales entre la multitud de factores involucrados en los fenómenos que determinan la actividad genética [43, 44]. A finales de la primera mitad del siglo XX, el embriólogo Conrad H. Waddington encaminó su trabajo en este sentido al concebir una teoría del desarrollo basada en las interrelaciones genéticas como un sistema dinámico multidimensional. Esta teoría fue posteriormente ampliada y formalizada y actualmente es el sustento teórico para la comprensión del desarrollo y la diferenciación celular desde el campo de la biología de sistemas [45–48].

En el campo de biología de sistemas, el desarrollo y la diferenciación celular son considerados procesos que resultan de la interacción entre los elementos que constituyen a un organismo. Entre los elementos considerados se incluyen factores intracelulares como genes, proteínas y metabolitos, y factores ambientales como moléculas de comunicación celular y elementos disueltos en el medio. Las relaciones entre estos elementos se abstraen bajo la forma de una red de regulación genética, que es un sistema dinámico que determina causalmente la actividad genética a partir de interacciones regulatorias entre los elementos que influyen al compor-

tamiento celular. El propósito de considerar a la célula un sistema dinámico es entender el comportamiento de éste en el tiempo analizándolo matemáticamente. En el centro de esta aproximación teórica está el concepto de estado del sistema, que se refiere a un conjunto mínimo de variables descriptivas del sistema y corresponde a un tipo de memoria que se actualiza constantemente. En este sentido, las redes de regulación genética determinan el comportamiento celular dependiendo de los elementos presentes en la célula en un tiempo dado (el estado del sistema) y las interacciones regulatorias que existen entre ellos. Desde un punto de vista muy general, un sistema puede mantenerse en el mismo estado en el que se encuentra o cambiar de estado. Cuando un sistema alcanza un estado en el cual no cambia en el tiempo se conoce como un estado estable o atractor y la lógica regulatoria del mismo sistema lo hace permanecer ahí. Algunos sistemas pueden tener diferentes atractores, éstos son conocidos como sistemas multiestables. Desde la perspectiva sistémica, los diferentes tipos celulares de un organismo corresponden a los atractores determinados por la red de regulación subyacente y el desarrollo es el movimiento de las células a través de los atractores del sistema.

Waddington creó la popular imagen del paisaje epigenético (Figura 1.2) como una representación diagramática de su idea de una concatenación de procesos unidos en una red guiando al desarrollo, haciendo que las células alcancen estados estables claramente distintos [42]. El estudio del desarrollo como un sistema dinámico ha tenido importantes avances desde que Waddington introdujo sus ideas. Stuart Kauffman formalizó la idea de una red global de regulación genética determinando la diferenciación celular como un sistema multiestable de atractores [49]. Los tipos celulares han sido asociados a atractores con los niveles de expresión genética como estado del sistema [45] y se han construido módulos de redes de regulación genética empíricamente, simulando su dinámica y usándolas para entender procesos de diferenciación. [50, 51].

Partiendo del planteamiento anterior, para entender el fenómeno de la troncalidad es necesario abordar su estudio desde la perspectiva del paisaje epigenético. Entonces, las células troncales representarían atractores dentro del paisaje epigenético con propiedades que les permiten mantener tanto la capacidad de auto-renovarse como el potencial de diferenciación celular. Existen múltiples aproximaciones al estudio de las células troncales desde perspectivas sistémicas. Dentro de éstas, hay investigaciones teóricas que buscan los principios básicos necesarios para mantener la troncalidad [21, 52–55] y también modelos de redes de regulación genética basados en evidencia empírica para simular computacionalmente el comportamiento de células troncales embrionarias [18, 56, 57].

Las indagaciones teóricas han hecho propuestas interesantes sobre atributos sistémicos asociados a la troncalidad. Una de estas propuestas plantea la necesidad del atractor troncal de mantener un balance entre robustez y plasticidad, que permitiría a la célula troncal mantenerse en el mismo estado al duplicarse o tener la libertad de explorar diferentes estados para diferenciarse en tipos celulares más especializados [21, 54]. Una manera de que el atractor tenga esas características sería que se encontrara en un estado de equilibrio dinámico que permitiera fluctuaciones en los niveles de expresión genética. Esto significa que cuando el sistema se encuentra en ese atractor, no se queda estático en un único estado sino que el mismo atractor está formado por diferentes estados entre los que oscila el sistema. En particular, una dinámica como esta ha sido asociada a un atractor

caótico, que le permitiría al sistema explorar el espacio de estados mientras que mantiene la robustez de seguir siendo un atractor determinado por la lógica regulatoria subyacente. Esto significa también que la heterogeneidad transcripcional no sólo sería consecuencia de la estocasticidad de los procesos intracelulares, sino que está determinada por el sistema de regulación y como tal tiene consecuencias funcionales [21, 28, 53, 57]. Siguiendo esta idea, la dinámica caótica de las células troncales es estabilizada durante el desarrollo gracias al aumento en el número de células que conlleva a que interacciones intercelulares impongan restricciones en la dinámica de células diferenciadas y por lo tanto éstas adquieren un perfil transcripcional estable y homogéneo [52]. Estos enfoques también han abordado el fenómeno de la reprogramación celular, entendida en este caso como la posibilidad de regresar al atractor troncal desde tipos celulares diferenciados. Desde este punto de vista, partiendo de un estado diferenciado con una dinámica estable, los experimentos de reprogramación celular desestabilizan la dinámica del sistema ocasionando que la célula regrese a una dinámica caótica y por lo tanto a un estado troncal [21, 58]. Estos análisis parten de primeros principios y logran dar explicaciones fundamentales a fenómenos asociados a la troncalidad como la heterogeneidad transcripcional y la reprogramación celular.

Los modelos empíricos de células troncales embrionarias se basan en su mayoría en la dinámica de una red mínima de regulación entre Oct4, Sox2 y Nanog [18]. Como se mencionó antes, estos tres factores de transcripción son los genes marcadores de células troncales embrionarias. Estudiando sus interacciones, se ha descubierto que entre ellos existen asas de retroalimentación positiva que refuerzan mutuamente su expresión [59]. Además del módulo de Oct4, Sox2 y Nanog, la red regulatoria se ha ampliado para incluir muchos otros genes que son regulados por estos factores de transcripción y que se sabe que tienen efectos en la troncalidad [60]. A partir de estas evidencias experimentales, se han construido modelos dinámicos de redes mínimas que recuperan el comportamiento observado de regeneración y diferenciación a partir de estímulos [18, 56]. Estos modelos sirven para entender los mecanismos de regulación genética involucrados en la troncalidad y se siguen actualizando a la luz de nuevas evidencias experimentales.

### 1.3. Una aproximación integrativa a la troncalidad

Los modelos dinámicos de redes de regulación de la expresión genética permiten simular la dinámica transcripcional en procesos de diferenciación celular. Sin embargo, como se mencionó antes, los fenómenos de la troncalidad y de la diferenciación celular no sólo implican cambios transcripcionales, sino que están asociados a cambios en el perfil epigenómico y la estructura global de la cromatina [36, 61–63]. Las modificaciones de la cromatina conllevan a mecanismos de regulación de la expresión genética que no han sido incluidos explícitamente en los modelos matemáticos dinámicos de regulación genética [48, 64]. Las modificaciones epigenéticas durante la diferenciación celular hacen que diferentes secciones del genoma sean accesibles o inaccesibles a factores de transcripción y por lo tanto implican cambios en la estructura global de la red de regulación [34, 63, 65]. La no incorporación de factores epigenómicos a los modelos teóricos de diferenciación

celular ha ocasionado un distanciamiento entre los ámbitos de biología de sistemas dinámicos y perspectivas más cercanas a la biología molecular, que describen cambios epigenómicos asociados a la diferenciación celular [47]. Lograr una vinculación entre las perspectivas dinámicas y epigenómicas de la troncalidad sería un avance importante para lograr un mejor entendimiento del desarrollo biológico y la diferenciación celular.

A partir de estas evidencias, busqué ahondar el conocimiento sobre la troncalidad, y en general sobre el desarrollo y diferenciación celular, realizando análisis estructurales a partir de datos epigenómicos e interpretándolos desde la perspectiva dinámica del paisaje epigenético. A través de este trabajo integrativo, pude inferir las interrelaciones entre el perfil epigenómico y la dinámica transcripcional, así como establecer una posible ruta teórica para unir las perspectivas del análisis dinámico de la diferenciación sintetizado en el paisaje epigenético con perspectivas estructurales de cambios epigenómicos en la cromatina asociados a la diferenciación celular.

## 2 Planteamiento del problema y objetivos

### Planteamiento del problema

Entender los fundamentos de la troncalidad es un reto pendiente. El concepto de la troncalidad es de especial interés debido a su íntima relación con el desarrollo de organismos multicelulares y por sus implicaciones para el desarrollo biomédico en fenómenos como el cáncer y la medicina regenerativa. El estudio de la troncalidad se ha abordado desde diferentes perspectivas, destacando características funcionales, transcripcionales y epigenómicas, distintivas de células troncales. Un reto pendiente para alcanzar un mejor entendimiento de la troncalidad es integrar las diferentes evidencias asociadas a ésta bajo una misma teoría general del desarrollo. Con esta tesis busco contribuir a este propósito a través del análisis de propiedades sistémicas de la red de regulación subyacente a la troncalidad, interpretándolas desde del marco conceptual del paisaje epigenético de atractores como modelo del desarrollo biológico y diferenciación celular.

### Objetivo general

Entender a la troncalidad como un fenómeno del desarrollo dentro del paisaje epigenético de atractores, integrando las implicaciones estructurales y dinámicas que tienen sobre la red de regulación genética los cambios epigenómicos que se dan durante el desarrollo y diferenciación celular.

### Objetivos particulares

1. Entender los fundamentos conceptuales y matemáticos de la teoría del paisaje epigenético de atractores como herramienta para modelar el desarrollo biológico.
2. Identificar propiedades sistémicas subyacentes a la troncalidad mediante el análisis e integración de datos de escala genómica, siguiendo una perspectiva de biología de sistemas complejos.
3. Encontrar características estructurales de redes de regulación genética específicas de tejido que distingan entre células troncales y tipos celulares diferenciados, analizando sus efectos en la dinámica del sistema.

4. Analizar propiedades funcionales de genes con diferencias en su nivel de tolerancia a las mutaciones para explorar la dependencia de contexto en la definición de función genética.
5. Integrar los cambios estructurales en redes de regulación inferidos a partir del análisis de datos epigenómicos a la teoría de modelado del paisaje epigenético de atractores.

### 3 Modelado del paisaje epigenético de atractores

A view of the biological world that reduces cause to discrete genetic and environmental forces reduces living beings to infinitely thin membranes resonating to signals from within or without but lacking the substance to generate signals of their own.

---

Susan Oyama,  
*The Ontogeny of information*, 2001

El desarrollo de un organismo multicelular es un proceso complejo en el que intervienen un sinnúmero de factores internos y externos [66]. Una pregunta fascinante que hoy en día sigue atrayendo atención es, ¿dónde se encuentra la información que guía al desarrollo biológico, haciendo que sea un proceso robusto y ordenado? Durante las décadas que siguieron al descubrimiento de la estructura del DNA, la preeminencia de la biología molecular y un enfoque reduccionista de la biología hicieron que la carga explicativa de los fenómenos observados en los seres vivos fuera atribuida a la acción de genes [44]. Sin embargo, esta visión genocéntrica es insuficiente para explicar procesos de desarrollo y diferenciación celular en los que a partir de la misma información genética se adquieren diferentes fenotipos [48, 67]. La información que guía a estos procesos no puede ser atribuida a una entidad en particular, sino que emerge de las interrelaciones que tienen los elementos involucrados en el proceso [68, 69].

Conrad H. Waddington propuso el modelo conceptual del paisaje epigenético en un intento por comprender al desarrollo de los organismos como un sistema dinámico guiado por las interrelaciones entre los elementos subyacentes a éste, en particular la acción de genes y su relación con el ambiente [42]. En su planteamiento original, el paisaje epigenético ilustra el proceso de diferenciación celular como el descenso de una esfera a través de una valle marcado por bifurcaciones (Figura 1.2). En esta imagen, la esfera representa a las células, que durante el desarrollo toman diferentes trayectorias hasta llegar a destinos separados por elevaciones del terreno, correspondiendo a los diferentes tipos celulares. Tal vez la contribución más interesante de este planteamiento es que el paisaje mismo por el que las células se mueven está determinado por un sistema complejo de interacciones que forman las crestas y valles que guían al desarrollo. En este sentido, el desarrollo



biológico es un proceso que está siendo continuamente determinado a partir de las interacciones que se dan entre los elementos que forman las células y su entorno, el paisaje epigenético no está dado previamente sino que es resultado de el mismo proceso dinámico del desarrollo.

Posteriormente a la publicación de las ideas de Waddington, el paisaje epigenético ha sido sustentado en el marco teórico de la teoría de sistemas dinámicos. Esta aproximación plantea que de manera subyacente al desarrollo existen redes de regulación genética que imponen restricciones al comportamiento de las células de un organismo, guiando de este modo la diferenciación celular [44]. El análisis computacional de la dinámica de redes de regulación genética construidas a partir de datos experimentales permite el estudio teórico de procesos de diferenciación celular.

### 3.1. Modelado computacional del paisaje epigenético

Una red de regulación genética es una representación formal de conocimiento fragmentado de mecanismos moleculares de regulación involucrados en procesos de desarrollo que es integrado usando herramientas de la teoría de sistemas dinámicos. La teoría de sistemas dinámicos busca entender cómo un sistema, entendido como un conjunto de elementos interrelacionados, se comportará en el tiempo al analizarlo a través de una formalización matemática. Existen diferentes maneras por las que un sistema puede ser descrito matemáticamente, por ejemplo sistemas de ecuaciones diferenciales ordinarias o parciales o lógica Booleana. El método elegido para modelar al sistema depende del nivel de detalle que se busca en la caracterización, así como de la cantidad de información disponible. Un concepto central en la perspectiva de sistemas dinámicos es el de estado del sistema, que se refiere a un conjunto mínimo de variables descriptivas del sistema, estas variables representan a los elementos que forman parte del sistema. El sistema puede ser descrito en cualquier tiempo por su estado y su comportamiento es causalmente determinado por el estado en el que se encuentra y las interacciones regulatorias que existen entre sus elementos. De esta manera, el estado del sistema es un tipo de memoria que se actualiza constantemente. Desde un punto de vista muy general, el sistema tiene dos comportamientos básicos: mantenerse en el mismo estado o cambiar de estado. Cuando un sistema alcanza un estado en el cual las reglas regulatorias hacen que no cambie en el tiempo, se dice que ha alcanzado un estado estable o un atractor del sistema. Algunos sistemas pueden tener varios atractores, estos sistemas se conocen como sistemas multiestables. La existencia de multiestabilidad se debe a que la manera en que los elementos que conforman estos sistemas están relacionados les permite existir en diferentes combinaciones sin que se produzcan cambios en su estado.

Para entender la idea de sistemas multiestables usaré como ejemplo un sistema bidimensional, esto es un sistema conformado por dos elementos. Todos los posibles estados de este sistema pueden ser representados en un plano Cartesiano, cada eje representando a uno de los elementos o variables del sistema. Esta representación gráfica es conocida como el espacio de estados del sistema. La dinámica del sistema es representada como trayectorias a través del espacio de estados. Considerando

que los elementos de este sistema promueven su propia producción y inhiben la del otro con la misma intensidad, el sistema puede alcanzar tres atractores. Estos son: la presencia de un elemento y la ausencia del otro y un estado en el que los dos elementos se encuentran en igual proporción. De esta forma, un sistema multiestable puede ser caracterizado por una red de interacciones tan sencilla como esta. La idea del espacio de estados puede ser extendida a sistemas de más dimensiones, aunque no podemos concebir visualmente sistemas de más de tres dimensiones.

Volviendo a la biología de sistemas, si una célula es el sistema que se va a estudiar, los elementos que la constituyen son los genes, proteínas, metabolitos y factores ambientales que tengan alguna incidencia en su comportamiento. Cada elemento o variable del sistema puede ser descrito en el tiempo por su nivel de actividad (la interpretación de esto depende de qué tipo de elemento se trate y de la formalización matemática utilizada) y las interacciones entre ellos conforman la red de regulación que subyace al comportamiento celular. De acuerdo a la teoría del paisaje epigenético, el conjunto de atractores de este sistema complejo multidimensional corresponden a los diferentes tipos celulares alcanzados durante el desarrollo. Considerando que el genoma de un organismo multicelular contiene miles de genes, la cantidad de posibles estados del sistema es enorme. Sin embargo, la manera en que estos elementos están interrelacionados restringe el número de estados estables en los que puede encontrarse el sistema. En este sentido, una trayectoria del desarrollo es el resultado dinámico de un conjunto de restricciones regulatorias.

Como se mencionó antes, desde esta perspectiva el desarrollo es el paso de las células del organismo a través del paisaje epigenético. Entonces, para analizarlo no es suficiente describir únicamente los estados estables del sistema, sino que se requiere poder estudiar las transiciones entre estados del sistema que caracterizan al desarrollo. La teoría de sistemas dinámicos tiene herramientas conceptuales para estudiar las transiciones entre estados estables de un sistema, dadas las condiciones dinámicas necesarias. Estas transiciones se pueden dar gracias a señales endógenas o exógenas que hacen que el sistema cambie de estado, o bien por fluctuaciones estocásticas en las variables del sistema que lo hacen llegar a otro estado estable. Al introducir estocasticidad a los modelos dinámicos del desarrollo se puede cuantificar la probabilidad de transitar entre los diferentes estados estables. La estocasticidad permite caracterizar a los estados estables de acuerdo a su estabilidad relativa, que se interpreta como la probabilidad de salir de un atractor una vez que el sistema se encuentra ahí. La caracterización de los atractores por su estabilidad relativa permite hacer un vínculo directo entre los modelos de redes de regulación genética y el paisaje epigenético propuesto por Waddington. En términos generales, la estabilidad relativa de un atractor es proporcional a la profundidad de un valle en el paisaje epigenético. Esto significa que mientras más grande sea la estabilidad relativa de un estado, su posición en el paisaje epigenético se encontrará a mayor profundidad y por lo tanto es más difícil salir de ahí. En la Figura 3.1 represento gráficamente la idea general detrás del modelado del paisaje epigenético a partir de la teoría de sistemas dinámicos.

La complejidad involucrada en los sistemas de regulación genética del desarrollo y la imposibilidad de medir todos los parámetros que rigen las interacciones ha llevado al uso extendido de redes Booleanas para su modelado. El análisis de

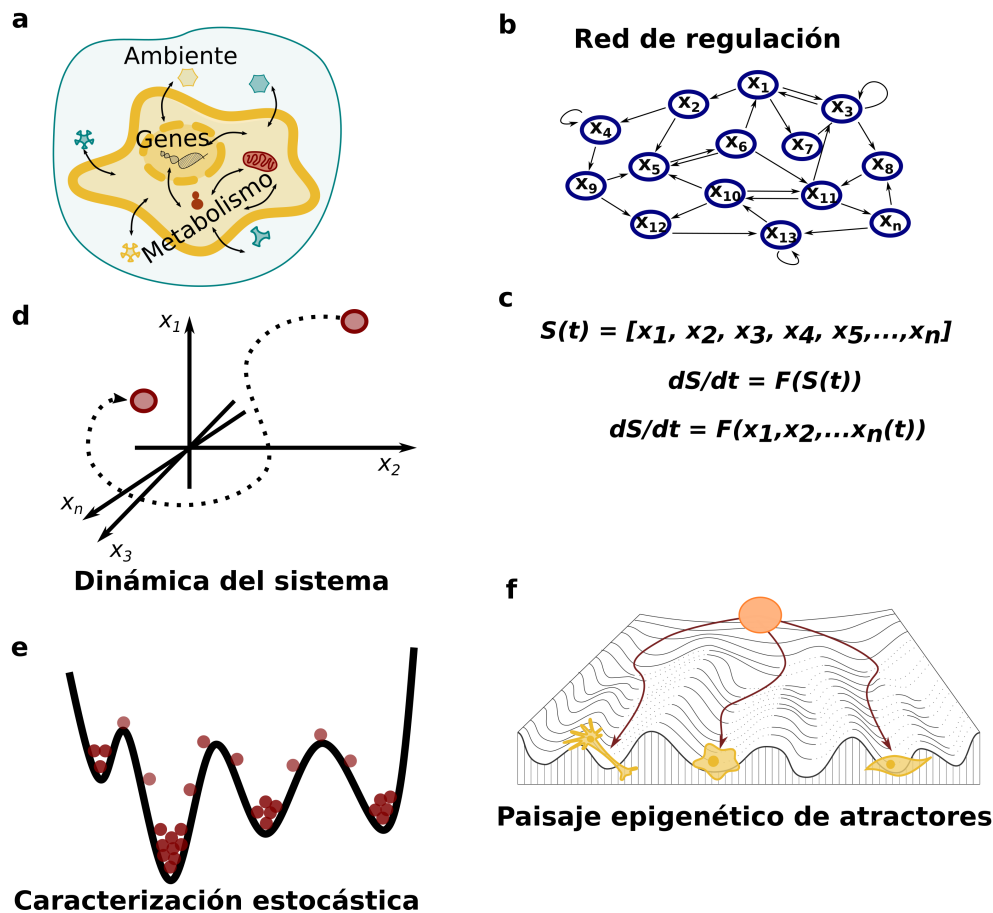


Figura 3.1: (a) La concepción sistémica del desarrollo plantea que el comportamiento de la célula emerge de las interrelaciones entre los elementos que la conforman y con el medio ambiente. (b) Su estudio parte de la abstracción de estas interacciones en forma de una red regulatoria y (c) su formalización matemática. (d) A partir de ésta se puede analizar el comportamiento dinámico del sistema y encontrar sus atractores. (e) Al introducir estocasticidad al sistema dinámico se pueden caracterizar la probabilidad de transitar entre estados y caracterizarlos en términos de estabilidad relativa. (f) De esta forma, se puede interpretar intuitivamente como un paisaje epigenético de atractores.

modelos Booleanos es relativamente sencillo al tener variables discretas de tiempo y estado. Los sistemas Booleanos tienen variables de estado binarias, esto quiere decir que los elementos sólo pueden estar en dos estados de actividad (encendido y apagado). El modelado de estos sistemas sigue los pasos esbozados anteriormente, en resumen se construye la red de regulación a partir de evidencias experimentales, se simula su dinámica computacionalmente para encontrar los atractores del sistema y finalmente se añade estocasticidad al modelo para encontrar el paisaje epigenético asociado. Para ahondar en estos conceptos, colaboré en un capítulo donde se explica detalladamente el método de modelado del paisaje epigenético usando redes Booleanas, este capítulo está incluido en el apéndice A.

## 3.2. Estudiando el cáncer a través del paisaje epigenético

El modelado de sistemas dinámicos permite abordar los fenómenos asociados al desarrollo desde una perspectiva integrativa que busca la causa de estos fenómenos no en un ente en particular, sino en las interacciones entre todos los elementos involucrados. El cáncer engloba un conjunto amplio de enfermedades que se caracterizan por el crecimiento celular descontrolado, esto la hace una enfermedad crónica degenerativa íntimamente relacionada al desarrollo [70]. El estudio tradicional del cáncer lo considera una enfermedad genética y por lo tanto su origen se explica por la acumulación de mutaciones en células somáticas, lo que ocasiona que se conviertan en células cancerígenas. A esta idea se le conoce como la teoría de mutaciones somáticas [71]. Este enfoque genocéntrico tiene limitaciones importantes al no tomar en cuenta los procesos complejos involucrados en la determinación del fenotipo celular [72]. Retomando lo que se ha mencionado antes, desde una perspectiva de sistemas la diferenciación celular está determinada por la dinámica de un sistema de regulación subyacente. Las transiciones entre tipos celulares son una propiedad fundamental del desarrollo de organismos multicelulares que suceden en ausencia de mutaciones genéticas. Desde este punto de vista, el cáncer es una enfermedad del desarrollo guiada por los mismos mecanismos involucrados en la diferenciación celular que normalmente producen la diversidad de tipos celulares presentes en el cuerpo. El estudio del cáncer mediante modelos de redes de regulación genética busca entender mecanismos genéricos que subyacen a la transformación oncogénica celular o tisular en diferentes tipos de cáncer.

Existen muchos estudios que han abordado el modelado del cáncer usando redes de regulación genética. Estos estudios buscan entender diferentes aspectos de la enfermedad que son difícilmente explicados a partir de mutaciones genéticas. En general, estos estudios se basan en el método general de modelado del paisaje epigenético expuesta en el capítulo anterior y analizan diferentes aspectos de la dinámica asociada a cierto tipo de cáncer. Partiendo de estas ideas, hice una revisión de cuatro investigaciones basadas en modelos de redes de regulación genética para el estudio del cáncer, esta revisión fue publicada como un capítulo y está incluida en el apéndice B.

Los estudios revisados en el capítulo se enfocan en explicar diferentes aspectos del cáncer abordándolos desde la perspectiva del paisaje epigenético de atractores, como se muestra en la Figura 3.2. El primer estudio analiza el fenómeno de heterogeneidad tumoral en cáncer gástrico. Proponen que esta heterogeneidad está asociada a que el fenotipo de cáncer está conformado por diferentes estados de la red, lo que permite que existan diversos tipos de células en los tumores sin que los diferentes fenotipos tengan que ser explicados por la aparición de mutaciones nuevas [73]. El segundo estudio busca encontrar los mecanismos moleculares mediante los cuales un tratamiento a base de ácido trans-retinóico (ATRA) funciona para revertir la leucemia promielocítica aguda, haciendo que se recupere una diferenciación granulocítica normal [74]. Al analizar la red regulatoria subyacente a la leucemia promielocítica aguda, encontraron que la actividad de ATRA influye en la forma del paisaje epigenético evitando que se alcance un atractor cancerígeno, explicando de esta forma el mecanismo de acción del tratamiento a nivel del

paisaje epigenético. El tercer estudio analiza un proceso conservado en diferentes cánceres epiteliales en los que la inflamación tisular antecede a la transición epitelio-mesénquima asociada a la aparición del cáncer. Este estudio encuentra un mecanismo conservado en el que la inflamación ocasiona la adquisición de un estado senescente que facilita la adquisición de un fenotipo mesenquimal [75]. Finalmente, el último estudio revisado analiza el papel que juega el gen  $TGF-\beta$  en la transición epitelio-mesénquima al modelar el comportamiento de una red de regulación genética de cáncer hepatocelular, concentrándose en el papel de este gen dentro de la red de regulación [76]. Analizar el papel de  $TGF-\beta$  al nivel de la red de regulación, permite entender su función en las transiciones de estado celular dentro del paisaje epigenético más allá de líneas causales de acción genética.

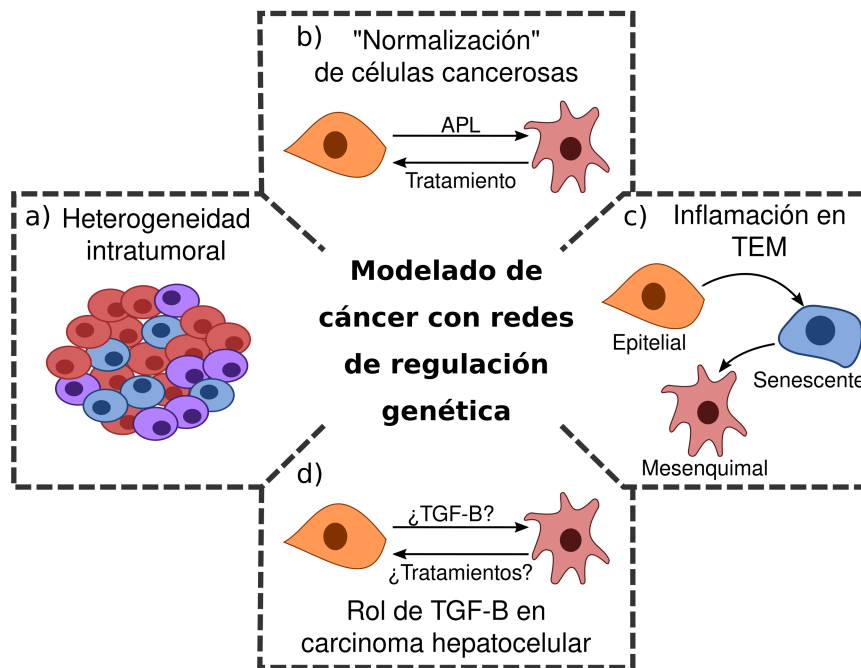


Figura 3.2: El estudio del cáncer mediante el uso de modelos de regulación genética permite abordar fenómenos asociados a la enfermedad que son difíciles de explicar por la teoría mutacional del cáncer. a) La heterogeneidad celular dentro de un tumor se puede explicar como diferentes estados de una red sin la necesidad de que estén asociadas a nuevas mutaciones. b) El análisis de la red subyacente a leucemia promielocítica aguda permite entender como un tratamiento clínico hace que las células cancerosas regresen a un fenotipo sano. c) El estudio de el efecto de la inflamación en tejidos epiteliales explica la transformación de este tejido en tejido mesenquimal pasando por un estadio senescente. d) Mediante el modelado de la red subyacente al carcinoma hepatocelular se puede entender el papel que tiene  $TGF-\beta$  en la aparición este cáncer.

### 3.3. La troncalidad y el cáncer

Al analizar las características del cáncer en el contexto de la troncalidad, es inevitable mencionar las similitudes que hay entre ambos fenómenos. La principal característica en común es la capacidad de reproducirse y generar más células, en las células troncales sanas de manera controlada y para el beneficio del organismo

mientras que en el cáncer es de manera descontrolada y afectando al organismo. Además de esta característica, comparten el alto grado de heterogeneidad transcripcional, la dependencia en el nicho en el que se encuentran, la capacidad de mantenerse con un metabolismo glucolítico, entre otras características [77]. Esto hace pensar que ambos fenómenos emergen de los mismos fundamentos que rigen la diferenciación celular y el desarrollo de organismos multicelulares. Sin embargo, la teoría de mutaciones somáticas, la teoría más extendida y aceptada sobre el origen del cáncer, plantea que los diferentes tipos de cáncer son causados porque una célula adquiere mutaciones en genes que ocasionan un cambio en su fenotipo, dando origen a una célula cancerígena que se multiplica y ocasiona cáncer [78, 79].

La teoría mutacional del cáncer es opuesta al entendimiento del desarrollo organísmico, en el que las células se diferencian formando diferentes órganos y tejidos sin la necesidad de que haya cambios genéticos. La teoría de mutaciones somáticas es incapaz de explicar la aparición del cáncer debido a que está basada en un mapeo lineal entre genotipo y fenotipo, además de que evidencias recientes han demostrado que las mutaciones somáticas no son causas suficientes ni necesarias para el cáncer [79]. Las mutaciones somáticas son causa insuficiente del cáncer debido a que tejidos sanos de adultos tienen en su genoma una acumulación de mutaciones en diferentes genes, incluidos oncogenes, sin que esas células adquieran un fenotipo maligno [80–82]. Las mutaciones somáticas tampoco son una causa necesaria para la aparición de cáncer ya que existen diferentes tipos de cánceres en los cuáles al estudiar el tejido canceroso no se encuentra ninguna mutación en oncogenes que pudiera explicar su transformación [79].

Volviendo a la idea de que el cáncer y la troncalidad surgen de los mismos mecanismos fundamentales del desarrollo y diferenciación celular y siguiendo la idea de que el desarrollo biológico puede ser entendido a través del paisaje epigenético, ¿cómo se puede estudiar el cáncer utilizando el paisaje epigenético? El fundamento teórico del paisaje epigenético es entender a la diferenciación celular como un proceso regido por un sistema multiestable subyacente en el que se alcanzan diferentes atractores que corresponden a los tipos celulares del organismo. Bajo esta idea, el sistema subyacente al desarrollo puede tener más atractores que los que son utilizados durante el desarrollo por tipos celulares, esto quiere decir que pueden existir estados estables sustentados por la dinámica del sistema que sin embargo son evitados durante el desarrollo normal de los organismos. Entonces, los diferentes tipos de cáncer vistos desde el paisaje epigenético son atractores inutilizados durante el desarrollo a los que las células del cuerpo pueden transitar por diferentes alteraciones en su estado, ya sean mutaciones genéticas, factores ambientales, mecanismos epigenéticos o fluctuaciones aleatorias [72, 78, 83]. Partiendo de esta idea, una aproximación al estudio del cáncer a partir del modelado del paisaje epigenético local de los tejidos a partir de los cuales se originan diferentes tipos de cáncer permite entender mejor los mecanismos que llevan a las células a adquirir un fenotipo canceroso sin tener que recurrir forzosamente a la explicación por mutaciones somáticas. Asimismo, entendiendo las características de los atractores de células troncales y asociándolas con su fenotipo permitiría encontrar coincidencias con los atractores de estados cancerosos que explicarían sus similitudes fenotípicas.

Los conceptos abordados en estos capítulos para entender la construcción empírica del paisaje epigenético y los beneficios de utilizarlo para el estudio del de-

sarrollo sustentan el entendimiento conceptual del desarrollo y la diferenciación celular que guía el resto de esta tesis. La metáfora del paisaje epigenético y su formalización a través de redes de regulación genética permiten estudiar al desarrollo como un proceso que emerge a partir de las interrelaciones entre los elementos celulares y ambientales subyacentes. Siguiendo esta idea, no existen entidades que guíen al desarrollo, sino que la información se encuentra en las interrelaciones de estos elementos y se genera en el proceso mismo del desarrollo [69].

## 4 Propiedades estructurales de la red de factores de transcripción de células troncales embrionarias

Understanding ontogeny thus becomes partly a matter of charting the shifts from one source of change (including intraorganismic processes) to another, as one interaction alters the developmental system in a way that provides transition to the next. Equally important are the means whereby stability is achieved.

---

Susan Oyama,  
*The Ontogeny of information*, 2001

En biología de sistemas, se asume que los comportamientos de una célula emergen de la dinámica de un sistema biomolecular subyacente [46, 84]. Este supuesto, unido al desarrollo de tecnologías moleculares de caracterización del perfil transcriptómico y epigenómico, han llevado al desarrollo de métodos para inferir redes de regulación específicas del contexto [63, 85, 86]. Estas redes buscan conocer empíricamente las interacciones regulatorias que suceden en un tipo celular o en un contexto ambiental específico. Al ser construidas a partir de una gran cantidad de datos moleculares, estas redes se presentan como una maraña de puntos o nodos unidos por aristas. En general, los nodos son genes y cuando se encuentran unidos por una arista significa que existe una interacción entre ellos, la interpretación precisa de estas interacciones depende de la forma y de los datos a partir de los cuáles son inferidas.

Una característica importante de este tipo de redes inferidas a partir de datos *-omicos* es que, a diferencia de las redes mencionadas en el capítulo anterior, éstas son estáticas, lo que significa que no se pueden utilizar para el modelado de sistemas dinámicos. Esto se debe a dos razones, primero, aunque las redes reflejan interacciones entre genes no se sabe la función que rige esta interacción. Es decir, una red puede indicar que el gen A influye en el comportamiento del gen B, sin embargo no se sabe si es una interacción de activación o inhibición. La segunda razón es que aunque se conociera la forma en que se da la interacción, estas redes son



demasiado grandes para simular su dinámica computacionalmente. A pesar de que no se pueden hacer análisis dinámicos a partir de estas redes, utilizando métodos de la teoría de redes, se puede analizar su estructura para obtener información sobre el fenómeno biológico al que están asociadas. A partir de análisis estructurales de redes se han descubierto que las redes biológicas tienen propiedades aparentemente universales de redes complejas, como una distribución de grado libre de escala y un alto grado de modularidad [84]. La capacidad de construir redes de regulación específicas de tejido permite explorar la forma en que estas propiedades varían en los diferentes contextos biológicos y de esta forma encontrar un vínculo entre las variaciones estructurales de la red y el comportamiento biológico asociado a éstas.

A partir de estas ideas, realicé un análisis estructural comparativo de redes de factores de transcripción de células troncales embrionarias y células diferenciadas. Estas redes fueron construidas a partir de datos de accesibilidad a la cromatina (DNase-seq) y publicadas en [86, 87]. La forma en que se infieren las interacciones de factores entre transcripción es buscando sitios de unión para un factor de transcripción en un sitio accesible a la cromatina dentro del promotor del factor de transcripción. Por ejemplo, supongamos que un factor de transcripción A tiene un sitio de unión en la región promotora del factor de transcripción B, si en un tipo celular este sitio de unión es accesible a la cromatina se registra una interacción dirigida de A a B, como se muestra en la Figura 4.1. Haciendo esto sucesivamente para todos los factores de transcripción para los que se conocen sus motivos de unión se puede construir una red global de factores de transcripción. Utilizando esta metodología se publicaron redes específicas de 41 tipos celulares de humano y 25 tipos celulares de ratón, incluyendo células troncales embrionarias y de diferentes tejidos adultos.

Una propiedad universal de las redes complejas, incluyendo a las redes biológicas, es que son robustas a errores, pero frágiles ante ataques dirigidos a sus nodos centrales. Esto es una consecuencia de su arquitectura, por la que tienen unos cuantos nodos altamente conectados mientras que la mayoría de los nodos tienen pocas conexiones [88]. Tomando en cuenta esta propiedad y con la disponibilidad de redes específicas de tipo celular, hice una caracterización estructural de estas redes basado en su la robustez ante ataques dirigidos y fallas aleatorias, como se muestra en la Figura 4.1. Este análisis mostró un comportamiento distintivo de las células troncales embrionarias, siendo significativamente más robustas a ataques que redes de tipos celulares diferenciados. Un análisis más profundo, basado en la descripción estructural de las redes y su comparación con modelos teóricos de redes, demostró que la robustez está asociada a una arquitectura de red menos ordenada, esto quiere decir que las conexiones entre los factores de transcripción tienen una distribución más amplia y homogénea. Los modelos de redes utilizados en la comparación fueron el modelo Erdős-Renyi (ER), en el que las aristas se asignan con igual probabilidad entre cualquier par de nodos resultando en una distribución de grado homogénea, y el modelo Barabási-Albert (BA), que construye las redes con un algoritmo de conexión preferencial que resulta en redes con una distribución de grado libre de escala. Una comparación entre la dinámica asociada a estos dos tipos de redes ha mostrado que la característica de redes tipo BA de ser libres de escala les permite acceder a una dinámica ordenada más fácilmente que redes tipo ER, en la que es más probable que tengan una dinámica caótica.

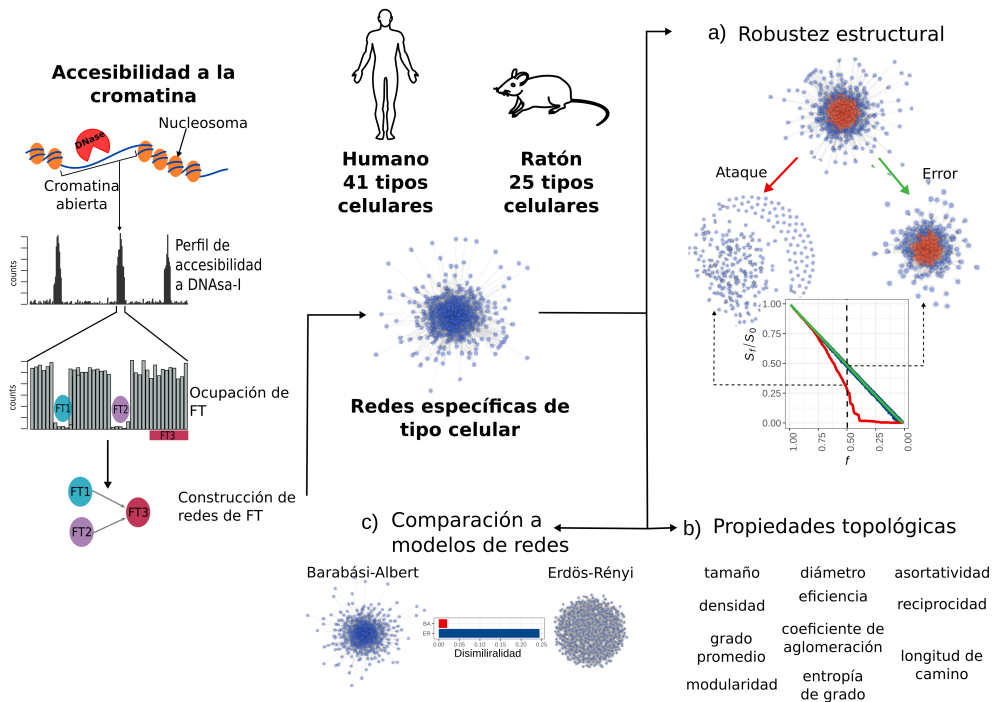


Figura 4.1: Análisis estructural de redes de factores de transcripción. En el panel izquierdo se muestra cómo se infieren las interacciones entre factores de transcripción a partir de datos de accesibilidad a la cromatina. Utilizando este método se construyeron redes de factores de transcripción específicas de tipos celulares de humano (41) y ratón (25). A partir de estas redes realicé un análisis estructural basado en la caracterización de la robustez estructural (a), la descripción estructural de las redes (b) y su comparación a modelos de redes aleatorias (c).

Entonces, la comparación de las redes de tipos celulares con modelos teóricos de redes aleatorias permite hacer inferencias sobre las implicaciones dinámicas de las diferencias estructurales observadas y relacionarlas con características biológicas.

El análisis estructural mostró que las redes de células troncales embrionarias se distinguen claramente de las de tipos celulares adultos por ser más robustas a los ataques por tener una estructura de red más homogénea y ser más cercana a una red tipo ER. Desde un punto de vista teórico, la arquitectura de red de las células troncales está asociada a una dinámica caótica que pensando en el paisaje epigenético les permitiría explorar una porción amplia del espacio de estados. Desde un punto de vista biológico, la estructura homogénea de la red está asociada a un perfil de cromatina más abierto y permisivo en las células troncales y la dinámica poco ordenada explicaría la heterogeneidad transcripcional. Este trabajo muestra como a partir de la integración de datos epigenómicos en redes y el análisis estructural de éstas se pueden hacer inferencias dinámicas con una interpretación biológica relevante. Estos resultados fueron publicados en un artículo que está incluido en el apéndice C.



## 5 Análisis de la dependencia del contexto en la definición de genes ‘esenciales’ en humano

It is not that the whole is more than the sum of its parts. It is that the properties of the parts cannot be understood except in their context in the whole. Parts do not have individual properties in some isolated sense, but only in the context in which they are found.

---

Richard Lewontin,  
*Biology as ideology*, 1991

Durante la segunda mitad del siglo XX, se han buscado explicaciones a los fenómenos biológicos preponderantemente en términos de genes como factores hereditarios que determinan la forma, la fisiología y el comportamiento de los organismos [89,90]. Partiendo de este supuesto, asociar una función determinada a un gen ha sido un objetivo fundamental en la biología evolutiva y del desarrollo [91]. Un método popular para estimar la importancia de un gen en el organismo ha sido medir su nivel de ‘esencialidad’. Un gen es considerado esencial si el organismo o la célula que lo contiene lo requiere para vivir [92]. En humanos, la determinación del nivel de esencialidad de un gen se ha realizado a través de dos vías. Un método es medirla a través de experimentos *in vitro*, en los que a partir de un cultivo celular se hace *knock-down* de cada uno de sus genes para probar su efecto en la viabilidad, la otra vía es a través de medidas poblacionales de variación genética *in vivo*, en los que la ausencia de mutaciones en un gen se asocia con su esencialidad. Cabe mencionar que estas definiciones de los contextos *in vitro* e *in vivo* son exclusivas de el estudio de genes esenciales en humano. De manera más general un estudio *in vivo* implicaría la experimentación con animales en laboratorio, sin embargo esto es imposible de realizarse en humanos. Hoy en día existen diferentes medidas de esencialidad de los genes del humano.

El supuesto de que la función y el papel que desempeña un gen en el organismo le es intrínseca al gen hace que estas medidas ignoren que los diferencias entre los contextos *in vitro* e *in vivo* en los que se mide la esencialidad. Sin embargo, esta idea ha sido cuestionada ampliamente arguyendo la dependencia de contexto en la función de los genes [69,93]. Siguiendo la idea planteada en los capítulos ante-

rios de que el comportamiento del organismo está determinado por un complejo sistema de interrelaciones subyacente, el papel de un gen depende de su situación dentro del sistema completo de interrelaciones. Este sistema de interacciones depende del contexto en el que sucede y por lo tanto las diferencias entre el contexto *in vivo* e *in vitro* son determinantes del nivel de esencialidad asignado a los genes en estos estudios. Como una aproximación a este debate, hice un análisis de genes esenciales, diferenciándolos dependiendo del contexto en el que fueron definidos (*i.e.* *in vitro* o *in vivo*). La hipótesis que guió este trabajo es que, dada la diferencia en el contexto de definición de la esencialidad, los genes obtenidos por cada enfoque tendrían características diferentes que estarían asociadas a las diferencias experimentales en las que fue medida la esencialidad.

Para hacer una comparación suficientemente amplia de las propiedades de los genes esenciales (intolerantes a las mutaciones de pérdida de función), integré bases de datos genéticas sobre aspectos funcionales, evolutivos, de expresión espacio-temporal y de asociación a diferentes enfermedades. Un primer resultado fue que irrespectivamente del contexto de definición, los genes intolerantes a las mutaciones tienen una posición más central en la red de interacción proteína-proteína que genes con mayor tolerancia a las mutaciones. Además de su posición en el interactoma, los genes intolerantes a las mutaciones se distinguen por ser más haploinsuficientes, tener una expresión más amplia en los órganos del cuerpo y codificar proteínas intrínsecamente desordenadas que carecen de estructuras secundarias o terciarias estables y cuya forma depende del contexto en el que se encuentran (Figura 5.1). Profundizando el análisis, encontré patrones contrastantes en cuanto a características funcionales dependiendo el contexto de definición de los genes intolerantes a las mutaciones. En particular, los genes esenciales definidos *in vivo* capturan genes asociados que tienen un origen evolutivo más reciente, están enriquecidos en proteínas involucradas en la comunicación y la diferenciación celular y están asociados a procesos del desarrollo y a funciones específicas del sistema nervioso central. En cambio, los genes esenciales definidos *in vitro* tienen un origen evolutivo más antiguo y están asociados a funciones críticas para la viabilidad celular como el metabolismo celular, la síntesis de proteínas y el ciclo celular (Figura 5.1).

Un resultado particularmente interesante de este análisis fue encontrar un subgrupo de genes con un comportamiento contrastante entre contextos. Estos genes son intolerantes a las mutaciones en análisis *in vivo* pero tolerantes a nivel *in vitro*, o viceversa. El grupo de genes intolerantes *in vivo* y tolerantes *in vitro* acentúa los patrones contrastantes entre contextos, teniendo un enriquecimiento en expresión en el sistema nervioso central, asociados a procesos del desarrollo y en su mayoría (80 %) siendo exclusivos de vertebrados. Además de esto, estos genes están enriquecidos en genes con asociación a enfermedades psiquiátricas y neurodegenerativas. Estos resultados refuerzan la idea de que los genes no pueden ser analizados de manera aislada, sino que tienen que entenderse como parte de un sistema complejo en el que su función depende del contexto y de las relaciones que mantengan con el resto del sistema. En particular, para medir la importancia de un gen en el humano es importante considerar su funcionamiento dentro del organismo completo.

En el contexto de la troncalidad, un resultado particularmente relevante es que lo que define a una célula troncal (autorenovación y diferenciación) está relacionado

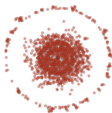



	Intolerantes		Tolerantes	
Restricciones asociadas a tolerancia mutacional	Agregadas Centrales 		Estructurales	Dispersas Periféricas 
	Funcionales		Funcionales	
	<ul style="list-style-type: none"> <li>△ . . . . . Haploinsuficiencia . . . . . ▽</li> <li>△ . . . . . Desorden intrínseco . . . . . ▽</li> <li>△ . . . . . Amplitud de expresión . . . . . ▽</li> <li>△ . . . . . Especificidad . . . . . △</li> </ul>		<ul style="list-style-type: none"> <li>▽</li> <li>▽</li> <li>▽</li> <li>△</li> </ul>	
	<div style="display: flex; justify-content: space-around;"> <div style="text-align: center;"> <i>in vitro</i> n=3028         </div> <div style="text-align: center;"> <i>in vivo</i> n=3139         </div> </div>		<div style="display: flex; justify-content: space-around;"> <div style="text-align: center;"> <i>in vitro</i> n=3028         </div> <div style="text-align: center;"> <i>in vivo</i> n=3139         </div> </div>	
Propiedades dependientes del contexto de definición	Conservación		Conservación evolutiva	
	<ul style="list-style-type: none"> <li>△</li> <li>▽</li> </ul>		<ul style="list-style-type: none"> <li>▽</li> <li>-</li> </ul>	
	Duplicación		Duplicación	
	<ul style="list-style-type: none"> <li>△</li> <li>▽</li> </ul>		<ul style="list-style-type: none"> <li>△</li> <li>-</li> </ul>	
	Cinasas		Cinasas	
	<ul style="list-style-type: none"> <li>△</li> <li>▽</li> </ul>		<ul style="list-style-type: none"> <li>△</li> <li>-</li> </ul>	
Receptores		Receptores		
<ul style="list-style-type: none"> <li>△</li> <li>▽</li> </ul>		<ul style="list-style-type: none"> <li>△</li> <li>-</li> </ul>		
FT		FT		
<ul style="list-style-type: none"> <li>-</li> <li>△</li> </ul>		<ul style="list-style-type: none"> <li>-</li> <li>▽</li> </ul>		
Cancer		Cancer		
<ul style="list-style-type: none"> <li>△</li> <li>▽</li> </ul>		<ul style="list-style-type: none"> <li>△</li> <li>▽</li> </ul>		
RBP		RBP		
<ul style="list-style-type: none"> <li>△</li> <li>▽</li> </ul>		<ul style="list-style-type: none"> <li>△</li> <li>▽</li> </ul>		
				
Expresión en tejidos		Expresión en tejidos		
<ul style="list-style-type: none"> <li>△</li> <li>▽</li> </ul>		<ul style="list-style-type: none"> <li>△</li> <li>▽</li> </ul>		
Genes de desarrollo		Genes de desarrollo		
<ul style="list-style-type: none"> <li>△</li> <li>▽</li> </ul>		<ul style="list-style-type: none"> <li>△</li> <li>▽</li> </ul>		

Figura 5.1: Resumen de las características asociadas a los genes intolerantes a las mutaciones. En el panel superior se muestran propiedades que distinguen a genes intolerantes a las mutaciones irrespectivamente del contexto de su definición. En el panel inferior se muestran características que presentan patrones contrastantes de enriquecimiento para genes intolerantes a las mutaciones dependiendo el contexto.

con funciones que no parecen importantes en estudios *in vitro*. El equilibrio entre autorenovación y diferenciación celular depende de mecanismos de transcripción genética que restringen la adquisición de un perfil transcripcional asociado a un linaje diferenciado. Las funciones celulares de regulación de la transcripción y los procesos de desarrollo requieren ser estudiados desde una perspectiva que considere al organismo en su totalidad para entender qué determina el comportamiento de una célula troncal.

Estos análisis están integrados en un artículo publicado en *Nucleic Acids Research Genomics and Bioinformatics*, incluido en el apéndice D.



## 6 El perfil epigenómico visto desde el paisaje epigenético de Waddington

The “informational” significance of any developmental influence, as we have seen, depends on the state of the entire developmental system, including genes, the rest of the phenotype, and relevant aspects of surround, and on the level and the type of analysis.

---

Susan Oyama,  
*The Ontogeny of information*, 2001

So there is clearly truth in the belief that the world can be broken up into independent parts. But that is not a universal direction for the study of all nature. A lot of nature, as we shall see, cannot be broken up into independent parts to be studied in isolation, and it is pure ideology to suppose that it can.

---

Richard Lewontin,  
*Biology as ideology*, 1991

El paisaje epigenético, planteado inicialmente por Waddington y posteriormente ampliado y formalizado durante la segunda mitad del siglo XX, es el modelo conceptual más adecuado para entender el desarrollo biológico como un proceso resultante de las interacciones entre los elementos biomoleculares que forman a los seres vivos [46, 47]. Sin embargo, el descubrimiento de la estructura del DNA y el desarrollo acelerado de técnicas para describir cada vez con mayor detalle la composición molecular de las células, ha llevado a una preeminencia del papel explicativo de entes moleculares para los fenómenos biológicos, ignorando los pro-



cesos complejos que los subyacen [89, 90]. En el campo de biología del desarrollo, esta tendencia ha llevado a que la carga explicativa de los procesos de desarrollo y diferenciación celular recaiga sobre las marcas epigenéticas que modifican la estructura de la cromatina y ejercen control sobre la expresión genética [31]. Asociado a este proceso, el término ‘epigenética’, propuesto inicialmente por Waddington dentro de su teoría del desarrollo como un sistema dinámico, ha pasado a ser un término amplio que abarca casi cualquier fenómeno molecular con efectos en el fenotipo diferente a cambios genéticos. Así sucede que a pesar de que actualmente epigenética es una palabra bastante popular, no siempre queda claro a lo que se refiere y en general su uso difiere a la definición pensada inicialmente por Waddington [94, 95]. Más allá del término usado, el enfoque en entes moleculares como causantes de los cambios de expresión genética ha hecho que la perspectiva propuesta por Waddington de entender al paisaje epigenético como el resultado dinámico de un sistema complejo de interacciones subyacentes no sea tomado en cuenta. Esto ha ocasionado un distanciamiento entre ramas de la biología afines al modelado teórico y ramas más cercanas a la descripción molecular de modificaciones epigenómicas para el estudio del desarrollo [47].

El desarrollo acelerado de técnicas de descripción molecular también ha ocasionado que los modelos del paisaje epigenético a partir de redes de regulación genética no se desarrollen con la suficiente velocidad para incorporar los nuevos conocimientos moleculares que se generan [48]. Una crítica que se ha hecho a los modelos de paisaje epigenético a partir de redes de regulación genética es que la lógica regulatoria con la que se construyen está basada en mecanismos de regulación tipo ‘switch’ como los que operan en bacterias. Sin embargo, en eucariontes y en particular en metazoarios existen otros mecanismos de regulación de la expresión genética basados en la cromatina que hacen más complicada su incorporación a formalismos matemáticos de sistemas dinámicos [48, 64, 96]. Establecer vínculos entre las perspectivas teórica y molecular de la epigenética es indispensable para alcanzar un mejor entendimiento de los procesos del desarrollo. En este capítulo hago una propuesta en este sentido, buscando incorporar el perfil epigenómico de una célula dentro del paisaje epigenético de atractores basándome en los resultados presentados en los capítulos 2 y 3 de esta tesis.

Como se mencionó antes, el perfil epigenómico de una célula o tipo celular representa una descripción estática de un nivel de regulación genética codificado en la cromatina [32]. Algo que es necesario recalcar es que las modificaciones epigenómicas son adquiridas durante el desarrollo y su presencia en la cromatina depende de la acción de enzimas y mecanismos intracelulares. Entonces, el perfil epigenómico de una célula debe ser considerado como un resultado del proceso que guía al desarrollo. Los modelos de redes genéticas de regulación tradicionalmente parten de la base de un sistema de regulación constante (al menos para cierto proceso del desarrollo) del cual emergen diferentes perfiles de expresión genética. En este sentido, la incorporación de modificaciones epigenómicas en los modelos de redes de regulación es un reto al representar cambios en la estructura misma de la red de regulación subyacente. Una posible alternativa a este problema es considerar al perfil epigenómico de una célula como una descripción del estado del sistema en un momento dado, equivalente a la descripción del estado en cuanto a perfil transcripcional. Siguiendo esta idea, redes de regulación específicas de contexto inferidas a partir del perfil epigenómico (como las analizadas en el apéndice

C) son también una descripción del estado del sistema. Sin embargo, estas redes representan a su vez un sistema dinámico que tiene asociada una dinámica propia. Lo que esto refleja es que durante el proceso del desarrollo, las células adquieren perfiles epigenómicos que conllevan a una reestructuración del sistema de interacciones que guía el proceso mismo de diferenciación celular. De esta forma, las modificaciones epigenómicas influyen en la dinámica misma del sistema subyacente, funcionando como un mecanismo de retroalimentación por medio del cual la dinámica del sistema de regulación se sincroniza con el estado de desarrollo.

En el caso particular de las células troncales, caracterizadas por tener un perfil epigenómico laxo que permite que haya más interacciones entre factores de transcripción. Su perfil epigenómico particular ocasiona que el sistema tenga una dinámica menos estable que el de células diferenciadas, el cuál le permite explorar más libremente el espacio de estados. Conforme el desarrollo avanza, las células adquieren modificaciones epigenéticas que integran diferentes señales ambientales e intracelulares y como resultado de esto su dinámica va siendo restringida en los diferentes tipos celulares adultos. Las modificaciones epigenómicas son entonces un nivel de regulación por medio del cual el proceso del desarrollo incide en su propia dinámica regulatoria. Vistas desde el paisaje epigenético las modificaciones epigenéticas influyen en la topografía del terreno generando valles y montañas regulando de esta forma la dinámica del sistema. Estas ideas las exploro con más detalle en un ensayo que incluyo en el apéndice E.



## 7 Discusión y conclusiones

Yo había visto demasiadas cosas  
poco claras como para estar  
contento. Sabía demasiado y no  
suficiente.

---

Celine, *Viaje al fin de la noche*,  
1932

The problem is to construct a  
third view, one that sees the entire  
world neither as indissoluble whole  
nor with the equally incorrect, but  
currently dominant, view that at  
every level the world is made up of  
bits and pieces that can be  
isolated and that have properties  
that can be studied in isolation.

---

Richard Lewontin,  
*Biology as ideology*, 1991

La troncalidad es un concepto central para entender el desarrollo y la diferenciación celular. Sin embargo, el entendimiento de la troncalidad ha resultado elusivo a análisis reduccionistas que buscan explicarla por la acción de unos cuantos entes moleculares [2, 3]. La troncalidad y el desarrollo de los organismos en general, son resultado de la dinámica de un sistema complejo de interrelaciones entre una multitud de factores. Esto dificulta la atribución simple de causa y efecto a algunos factores involucrados, obligando a ir más allá del estudio de genes o, en casos más recientes de marcas epigenéticas, como explicación del desarrollo y la diferenciación celular [69,97]. Esta tesis presenta una propuesta para el estudio de la troncalidad bajo la premisa de integración de análisis de datos empíricos con planteamientos teóricos sustentados en la teoría de sistemas dinámicos.

La teoría del paisaje epigenético, como representación de un sistema dinámico complejo subyacente al proceso del desarrollo, es una vía teórica para la comprensión de los mecanismos de diferenciación celular. La gran ventaja de esta teoría es su planteamiento de la emergencia de los procesos del desarrollo a partir de la dinámica de interacción entre la multitud de factores moleculares subyacentes. El tratamiento formal del paisaje epigenético está dado por el modelado compu-

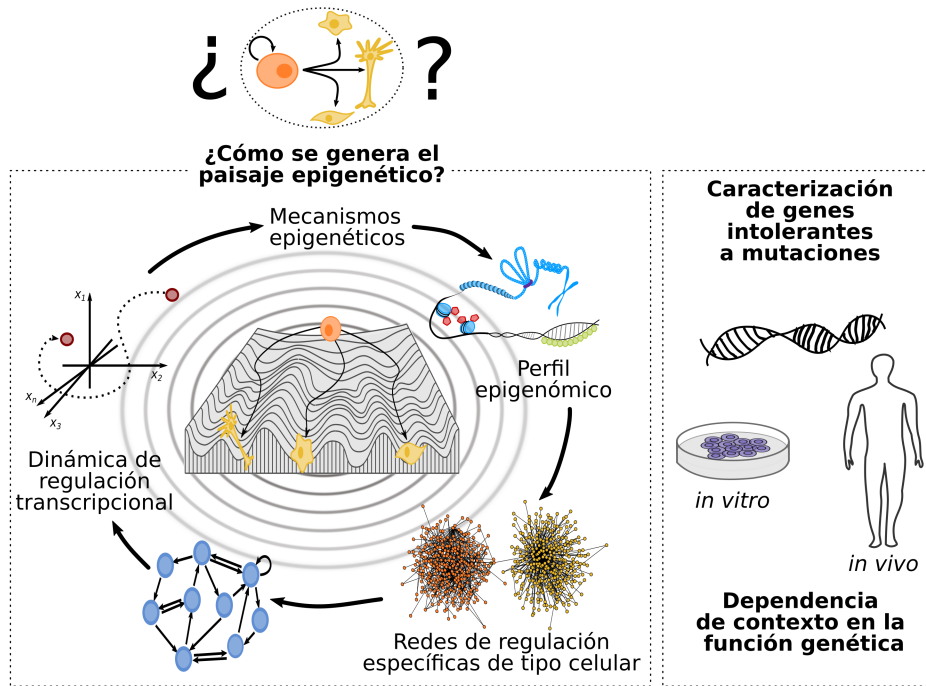


Figura 7.1: Resumen gráfico de mi trabajo de doctorado.

tacional de redes de regulación genética obtenidas a partir de datos empíricos. Sin embargo, el ritmo acelerado con el que han avanzado las técnicas moleculares de caracterización genómica y epigenómica ha dificultado la incorporación de nuevos conocimientos generados en estos campos a los modelos de redes dinámicas de regulación genética. Aún así, concebir a los organismos como un sistema dinámico complejo es la propuesta más adecuada para abordar el estudio del desarrollo y la diferenciación celular, ya que no le atribuye agencia a ningún elemento en particular sino a la interacción que se da entre todos los elementos del sistema [90].

Con este precepto, esta tesis comienza con una revisión de los métodos de modelado computacional del paisaje epigenético de atractores a partir de redes de regulación genética. Esta revisión examina los detalles técnicos del modelado del paisaje epigenético, así como sus ventajas como método de estudio del desarrollo. La revisión de los métodos de modelado de redes de regulación genética demuestra que el paisaje epigenético está sustentado en formalismos teóricos y evidencias empíricas y no es solamente una metáfora visual del desarrollo. La comprensión del sustento teórico detrás del paisaje epigenético permite también interpretar evidencias empíricas dentro de este marco conceptual de manera precisa. En un ejemplo práctico, el análisis de modelos de cáncer a través de redes de regulación genética permite ir más allá de la explicación mutacional como el origen del cáncer e interpretar fenómenos asociados a éste como la heterogeneidad tumoral, la influencia del entorno inflamatorio y el rol de algunos genes específicos en los procesos de decisión del destino celular. Asimismo, el tratamiento del cáncer por medio de redes de regulación genética abona a su entendimiento como una enfermedad asociada al desarrollo regida por los mismos principios que el desarrollo convencional de los organismos.

El marco teórico del paisaje epigenético permite avanzar en el estudio de fenómenos asociados al desarrollo, como la emergencia del cáncer y la troncalidad,

al concebirlos como diferentes procesos regidos por un mismo sistema de regulación subyacente. Esto permite analizarlos utilizando el mismo herramental teórico-computacional, el análisis de redes dinámicas de regulación genética. De este modo, fenómenos como la heterogeneidad transcripcional y la adquisición de potencial de desarrollo en células diferenciadas, relacionados tanto al cáncer como a las células troncales, pueden ser analizados como procesos determinados en el sistema de regulación en lugar de concebirlos como aspectos aislados y particulares de cada fenómeno.

Una vez planteadas las bases teóricas del paisaje epigenético, abordé el estudio de datos epigenómicos mediante el análisis estructural de redes de factores de transcripción. Este análisis me permitió encontrar inferencias sobre la dinámica asociada al estado de las células troncales embrionarias a partir de datos epigenómicos y análisis estructurales de redes específicas de tipo celular. Estos resultados muestran una manera mediante la cual se pueden interpretar datos epigenómicos como información del estado y la posición que ocupan las células en el paisaje epigenético. La conclusión principal del artículo presentado en esta sección es que las células troncales tienen una arquitectura de red más robusta y a grandes rasgos menos ordenada que la de células diferenciadas, como resultado de tener una organización de la cromatina menos restrictiva.

Esta arquitectura de red está asociada con una dinámica menos estable, con tendencia a ser caótica. La idea de que la troncalidad esté relacionada con una dinámica caótica ha sido planteada previamente [21], sin embargo este estudio da un sustento experimental a estas ideas y las relaciona con el perfil epigenómico de cromatina abierta característico de las células troncales. Las inferencias dinámicas obtenidas por estos análisis también dan un sustento teórico al fenómeno de heterogeneidad transcripcional característico de las células troncales. Este estudio representa una vía para vincular evidencias epigenómicas con planteamientos teóricos de sistemas dinámicos del desarrollo, buscando tender puentes entre aproximaciones de descripción molecular de redes globales de regulación celular con modelos teóricos del desarrollo.

El estudio de la función de los genes ha sido uno de los objetivos principales de la biología molecular y del desarrollo. Una perspectiva reduccionista en biología ha llevado a pensar que existe una relación uno a uno entre gen y función, en la que el gen tiene una función inherente irrespectivamente del contexto en el que se encuentre [89]. La determinación de genes intolerantes a las mutaciones como genes esenciales, cuya función es indispensable para la supervivencia del organismo, sigue esta visión reduccionista al concebir a los genes aislados del sistema en el que actúan.

En el cuarto capítulo de esta tesis presento un análisis crítico comparando genes esenciales en humano dependiendo del contexto en el que fueron definidos como tal: *in vivo* o *in vitro*. Este análisis demuestra que dependiendo del contexto de definición, las listas de genes esenciales cambian y también las características funcionales de estos genes. En particular, se demuestra que estudios *in vitro*, basados en cultivos celulares, son incapaces de detectar la importancia de genes del desarrollo. Esto lleva a considerar que la función de los genes depende de las relaciones que tiene con el sistema en general y como tal no pueden ser estudiados en aislamiento de éste. Considerando estos resultados en cuanto al estudio de la troncalidad, la dependencia de contexto en la función de los genes abona a la idea

de que la troncalidad no puede ser atribuida a la acción de unos genes sino que es resultado del estado en el que se encuentre el sistema de regulación global que permite a las células alcanzar el estado troncal.

Entender la troncalidad y en general la diferenciación celular requiere de la integración de la mayor cantidad de información disponible sobre estos procesos bajo un marco conceptual suficientemente amplio. El impresionante desarrollo de las capacidades de descripción molecular guiado por técnicas moleculares y de secuenciación, le ha dado a estos acercamientos un rol central como explicación de fenómenos biológicos. Sin embargo, la caracterización molecular de las células no es suficiente para entender los mecanismos que gobiernan el proceso de diferenciación. La comprensión de los principios básicos que guían la diferenciación celular requiere de herramientas conceptuales que permitan entender a la información epigenómica dentro de una perspectiva sistémica. Esta integración se puede lograr incorporando información obtenida a partir del perfil epigenómico a la idea del desarrollo como un sistema dinámico complejo, como lo planteó originalmente Waddington. Retomando lo expuesto en el último capítulo, mi propuesta es que las modificaciones epigenómicas actúan como un asa de retroalimentación entre el estado del sistema y su red de regulación. Con esta idea se le da una comprensión sistémica a los perfiles epigenómicos al reconocer que son determinados por un sistema subyacente y por lo tanto son una característica del estado del sistema. Por el otro lado, al reconocer que el perfil epigenómico ejerce una influencia sobre la topología de la red de regulación, se le reconoce como un nuevo nivel de control por medio del cual se puede afectar la dinámica y estabilidad del sistema.

La troncalidad es un concepto trascendental en la biología del desarrollo. Su comprensión está unida al mapeo entre genotipo y fenotipo y a la posibilidad de que existan diferentes tipos celulares a partir de la misma información genética. La caracterización molecular de células troncales ha demostrado que éstas tienen un perfil epigenómico distintivo que de alguna manera determina su comportamiento excepcional. Para alcanzar una mejor comprensión de la troncalidad es necesario integrar estas evidencias empíricas dentro del marco teórico del paisaje epigenético de forma que la troncalidad se entienda como una particularidad dentro del mismo paisaje que determina el desarrollo a partir de las interacciones entre los elementos que forman a los seres vivos.

## 8 Perspectivas a futuro

Science is impelled by two main factors, technological advance and a guiding vision (overview). A properly balanced relationship between the two is key to the successful development of a science: without the proper technological advances the road ahead is blocked. Without a guiding vision there is no road ahead; the science becomes an engineering discipline, concerned with temporal practical problems.

---

Carl Woese, 2004

Aventurarse a hacer predicciones sobre el rumbo que tomará una disciplina siempre es arriesgado y lo más probable es que resulten equivocadas. Sin embargo, para concluir esta tesis expongo algunas ideas sobre el rumbo que creo que tendrá el estudio de la troncalidad.

Las técnicas de secuenciación y de caracterización molecular de las células siguen avanzando a ritmo acelerado, empujando a casi todas las ramas de la biología a su paso. El estudio del desarrollo y de la troncalidad ha sido y a mi parecer seguirá siendo guiado en gran parte por la caracterización cada vez más detallada de estos procesos. Las técnicas de caracterización de célula única han aumentado significativamente la resolución con la que se pueden describir los fenómenos biológicos. Un esfuerzo muy importante dentro de este campo es el Atlas Celular Humano, proyecto que busca generar mapas de referencia con la posición, función y características de todas las células del cuerpo humano [9].

Dos conceptos relacionados con la troncalidad que se abordan dentro del proyecto de Atlas Celular Humano son la definición de tipo celular y la plasticidad celular de los tipos celulares. Teniendo una descripción altamente detallada a nivel de célula única hace necesaria una revisión del concepto de tipo celular, cuestionando qué tanta variación puede haber dentro de un tipo celular y en qué punto se marca la división entre un tipo celular y otro. Asimismo, conocer la variación que existe en los diferentes tipos celulares y su capacidad de transitar entre diferentes estados hace forzoso pensar en la plasticidad celular y los procesos por los que esta se adquiere o se pierde. Estas ideas forzosamente impactarán al campo de la troncalidad ya que el concepto de célula troncal está basado en las ideas de tipo



celular y de plasticidad celular. Tener una descripción detallada de los diferentes tipos celulares presentes en el cuerpo humano, sus relaciones ontogenéticas y la capacidad de las células de transitar entre diferentes tipos celulares es una fuente de información sumamente valiosa y necesaria para entender mejor el desarrollo y diferenciación celular.

La acumulación de más y mejores datos descriptivos del estado celular va a guiar el desarrollo de la biología del desarrollo. Sin embargo, el entendimiento conceptual de la forma en que se integran los diferentes aspectos genéticos, transcripcionales, epigenéticos, microambientales, y demás, aún requiere de una teoría del desarrollo que sea lo suficientemente amplia para abarcarlos a todos. A mi parecer, el tratamiento del desarrollo a partir de la teoría de sistemas dinámicos complejos es el marco conceptual más adecuado para abordar este problema. El modelado dinámico de redes de regulación genética es una herramienta conceptual muy poderosa para entender la emergencia de diferentes estados celulares a partir de la misma información genética. Sin embargo, me parece que la distancia es cada vez mayor entre el nivel de detalle alcanzado por modelos teóricos y las descripciones moleculares experimentales. Encontrar la manera de acortar la distancia entre las aproximaciones experimentales y de modelado teórico del estudio del desarrollo es fundamental para que esta área avance.

Para entender la troncalidad, otra área que me parece que tendrá importantes avances en los próximos años es el estudio del origen de la multicelularidad y en específico de la aparición de los metazoarios. La troncalidad es una característica que presentan todos los organismos multicelulares. Hongos, plantas y animales todos tienen alguna forma de troncalidad que les permite generar un organismo multicelular con diferentes tipos de células con el mismo genoma. Estudiar la evolución de unicelularidad a multicelularidad ofrecerá ideas sobre los fundamentos de la troncalidad. Con la misma perspectiva evolutiva, también me parece muy interesante el estudio de la evolución de tipos celulares a través de un análisis comparativo entre diferentes linajes evolutivos. Dentro de este campo es muy interesante el estudio de la evolución de mecanismos epigenéticos para entender las consecuencias que estos tienen en la capacidad de los organismos para tener distintos tipos celulares y su plasticidad ante factores ambientales.

# Apéndice



# A Modeling the epigenetic landscape in plant development

Capítulo publicado como parte del libro *Computational Cell Biology*, editado por L. von Stechow y A. Santos Delgado y publicado en 2018 por Springer International Publishing.



# Chapter 17

## Modeling the Epigenetic Landscape in Plant Development

Jose Davila-Velderrain, Jose Luis Caldu-Primo,  
Juan Carlos Martinez-Garcia, and Elena R. Alvarez-Buylla

### Abstract

Computational mechanistic models enable a systems-level understanding of plant development by integrating available molecular experimental data and simulating their collective dynamical behavior. Boolean gene regulatory network dynamical models have been extensively used as a qualitative modeling framework for such purpose. More recently, network modeling protocols have been extended to model the epigenetic landscape associated with gene regulatory networks. In addition to understanding the concerted action of interconnected genes, epigenetic landscape models aim to uncover the patterns of cell state transition events that emerge under diverse genetic and environmental background conditions. In this chapter we present simple protocols that naturally extend gene regulatory network modeling and demonstrate their use in modeling plant developmental processes under the epigenetic landscape framework. We focus on conceptual clarity and practical implementation, providing directions to the corresponding technical literature. The protocols presented here can be applied to any well-characterized gene regulatory network in plants, animals, or human disease.

**Key words** Epigenetic landscape, Gene regulatory networks, Dynamical systems, Systems biology, Cell differentiation, Attractors, Morphogenesis, Development

---

## 1 Introduction

Twenty-first century biology is largely an interdisciplinary endeavor. The ever-increasing accumulation of molecular experimental data point toward the need of encompassing integrative approaches to better understand observations and to unravel underlying mechanisms [1, 2]. The development of a multicellular organism is a complex process involving the dynamical interplay of multiple genetic and nongenetic factors at different spatial and temporal scales [3]. Given the complexity of such intricate systems, computational modeling has been a natural approach for the study of developmental dynamics. In particular, systems-level mechanistic models have been instrumental to simulate and better understand

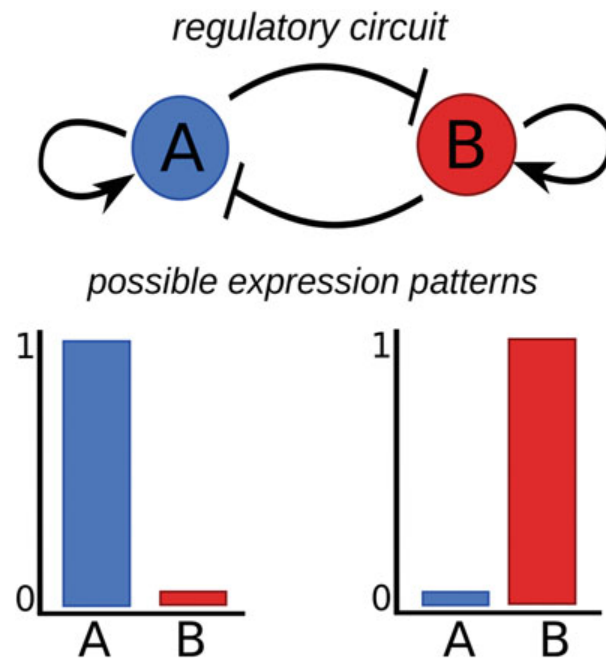
the forces driving cellular differentiation and morphogenesis in plant development [4].

One of the first attempts to understand the emergence of development as a dynamical system driven by complex interactions—as well as its association with genetics, the environment, and evolution—was the conceptual model of the Epigenetic Landscape (EL) proposed by C.H. Waddington [5]. In its original form, the EL depicts the development of a body part as a series of bifurcating trajectories undergone by balls rolling down a surface with hills and valleys, until reaching terminally differentiated states. The global shape of the landscape is assumed to be determined by underlying systems of complex genetic interactions. More recently, the EL has been grounded in the framework of gene regulatory networks (GRN), where cell states undergo differentiation transitions driven by the dynamics emerging from intracellular regulatory constraints [6]. In this chapter we focus on demonstrating how well established GRN modeling protocols can be extended in order to develop EL models for the study of cell differentiation and morphogenesis. We illustrate this approach on plant development examples as study systems.

GRN modeling has been successfully used for the theoretical study of developmental processes [4, 7–10]. We have previously presented general protocols for the implementation of GRN dynamical models and discussed their interpretation (*see* for example refs. 1 and 11). Given that GRN models are the basis to develop epigenetic landscape frameworks, we will include here a brief, nontechnical explanation of key concepts and methods for GRN modeling. We point the reader to our previous methods papers for a more in-depth exposition [7, 10, 11].

A GRN dynamical model is a formal representation of fragmentary knowledge about molecular regulatory mechanisms, which is integrated using the tools of dynamical systems theory [1, 2]. A GRN consists of a set of nodes and their mutual interactions (*see* **Note 1**). Both nodes and interactions are formalized in a set of rules that specify how the activity of each node affects the behavior of its interacting partners through time. In these models the activity of each node commonly represents the expression of a gene, and the goal of the model is to simulate the coordinated behavior of the complete gene expression profile. Because of the regulatory constraints specified in the dynamical rules, genes will display specific patterns of expression. For example, a regulatory constraint could simply indicate that an active gene can no longer remain expressed because its inhibiting partners become active (Fig. 1).

The power of the integrative GRN resides in its inherent ability to consider all such constraints in parallel and for each time step. The global consequence of the regulatory constraints



**Fig. 1** Regulatory constraints. A classic example of a regulatory circuit that imposes a regulatory constraint is the bistable switch shown in the Fig. A simple circuit is formed by two mutually repressing genes. This regulatory structure causes the system to have only three stable expression patterns: a trivial pattern in which none of the genes is expressed and two opposing behaviors in which either one of the two genes is expressed and the other gene is silent

is that only certain activity configurations (expression profiles) will be consistent with the regulatory logic and thus stably maintained in time. Such configurations are called attractors in the theory of dynamical systems and represent specific expression profiles in which the state of each gene does not change anymore (*see Note 2*).

Because of their robustness—(quasi-)stationary and stable— attractor states likely represent behaviors we would expect to observe in nature in terms of the gene activity characteristic of each cell type. This is a consequence of the information integrated in the GRN being experimentally grounded (for details, *see ref. 6 and 11*). For practical purposes, the attractors recovered using GRN dynamical models (*see Subheading 3* below) correspond to the cell states observed in real developmental systems and are empirically recovered by gene expression profiles.

The dynamical rules specifying the GRN can be modeled with different types of equations or procedural algorithms. Here, we will focus on models with discrete state and time variables, which are modeled with discrete map equations; we will focus on the simplest type of such models: the Boolean network model [12]. Previous studies have shown that such qualitative models capture the systemic behavior of GRNs underlying cell differentiation ([7, 9] and references therein). The capacity of Boolean models

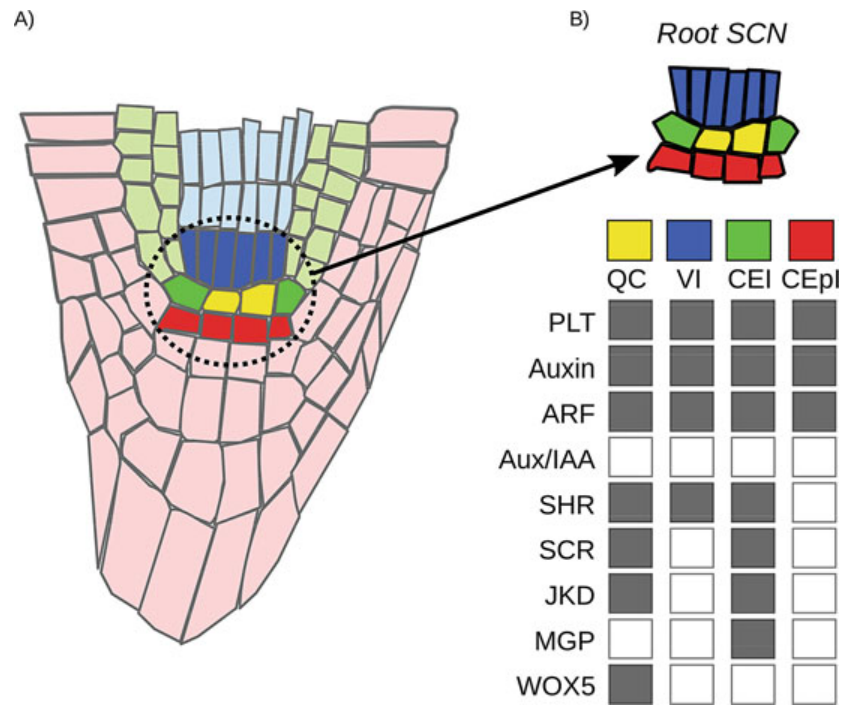
to recover systemic behaviors, despite their qualitative nature, evidences that biological behaviors are largely guided by the overall structure of the networks rather than the specific form of the kinetic functions [13].

In addition to successfully modeling cellular differentiation in diverse plant developmental processes [14–16], Boolean GRN models have also been used to operationalize the otherwise metaphorical model of the EL [6, 17–19]. In this context, as we shown below, Boolean GRNs are extended by the introduction of stochasticity in order to simulate cell state transitions (i.e., differentiation events), enabling the discovery of regular patterns of such transitions. These ideas were first applied in plant development in the study of the morphogenetic patterns observed during early flower development in *Arabidopsis* [17]. Based on this original study, as well as on more recent work proposing a general framework for modeling the EL associated with Boolean GRNs [6, 17], we recently implemented previously dispersed algorithms in a common modeling framework with the goal of facilitating the overall use of EL models [18]. In this chapter, we make use of some of these tools in order to provide a general protocol for modeling the EL associated with Boolean networks, a methodological standpoint that has been applied successfully in the study of several plant developmental processes [17, 20].

Given that the EL methods considered here constitute extensions to well-established protocols of GRN modeling and in order to provide a comprehensive overview, we also include a brief protocol for modeling Boolean GRNs. For all the examples we consider a well-studied developmental GRN: the root stem cell niche GRN (root SCN-GRN) of *Arabidopsis thaliana* [15].

The *A. thaliana* root stem cell niche (SCN) is part of the root apical meristem at the acropetal end of the primary root. The root SCN is well described at the anatomical level. It is formed by an inner core of four cells that rarely divide and constitute the quiescent center (QC), surrounded by three sets of initial cells that give rise to the different types of differentiated cells in the root: vascular initials (VI), cortex–endodermis initials (CEI), and columella–epidermis–lateral root cap initials (CEpI) (Fig. 2a) [15]. Some of the molecular components implicated in root SCN cellular patterning and maintenance have been uncovered and characterized. Three main pathways are involved. One component is the module of *SHORT-ROOT* (*SHR*) and its target gene *SCARECROW* (*SCR*); both genes regulate the transcription of several genes involved in the root cellular organization [21, 22]. Another component involves the *PLETHORA* (*PLT*) genes, which are transcribed in response to auxin accumulation [23]. Finally, *WUSCHEL RELATED HOMEODOMAIN 5* (*WOX5*) is a gene expressed exclusively in the QC, and it has been implicated in maintaining stable gene expression and keeping the distal SC





**Fig. 2** Simplified cellular pattern of the root stem cell niche. (a) Root tip of *Arabidopsis thaliana* colorized to show the corresponding cell types, highlighted is the root SCN. The four cell types within the SCN are: quiescent center (yellow), cortex–endodermis initials (green), vascular initials (blue), and columella–epidermis–lateral root cap initials (red). Root cells formed from the SCN are shadowed with the color of the corresponding SCN cell type. (b) Gene expression configuration in the cell types of the SCN. Gene activity is coded in squares: gray for active genes and white for inactive genes

undifferentiated [24]. Work in our group integrated the gene interactions recovered through molecular evidence and proposed a GRN that recovers the expected gene configurations of the different root SCN cell types [15] (Fig. 2b). In what follows, we will reanalyze such a network following the epigenetic landscape framework.

## 2 Materials

### 2.1 Data

The input to all the analyses described here is any previously specified Boolean GRN dynamical model (for example the model in [15]), which consists of a list of state variables (i.e., molecules of interest) and a corresponding set of Boolean logical propositions describing their regulatory interactions. For the root SCN example, we consider the experimental data in Table 4 and the Boolean propositions in Table 5. In addition, for validation and prediction purposes, observable wild-type cellular phenotypes and mutant (loss- and gain-of-function) gene activation configurations need to be previously defined from literature. We consider the gene expression profiles extracted from Azpeitia and collaborators [15].

## 2.2 Software

The R statistical programming environment ([www.R-project.org](http://www.R-project.org)) is used to implement all the analyses presented here, making use of the functions from the contributed package *BoolNet* [25], as well as those kept at <https://github.com/JoseDDesoj/Epigenetic-Attractors-Landscape-R>. The complete protocol for the analyses presented in this chapter can also be found in the Github folder mentioned above.

---

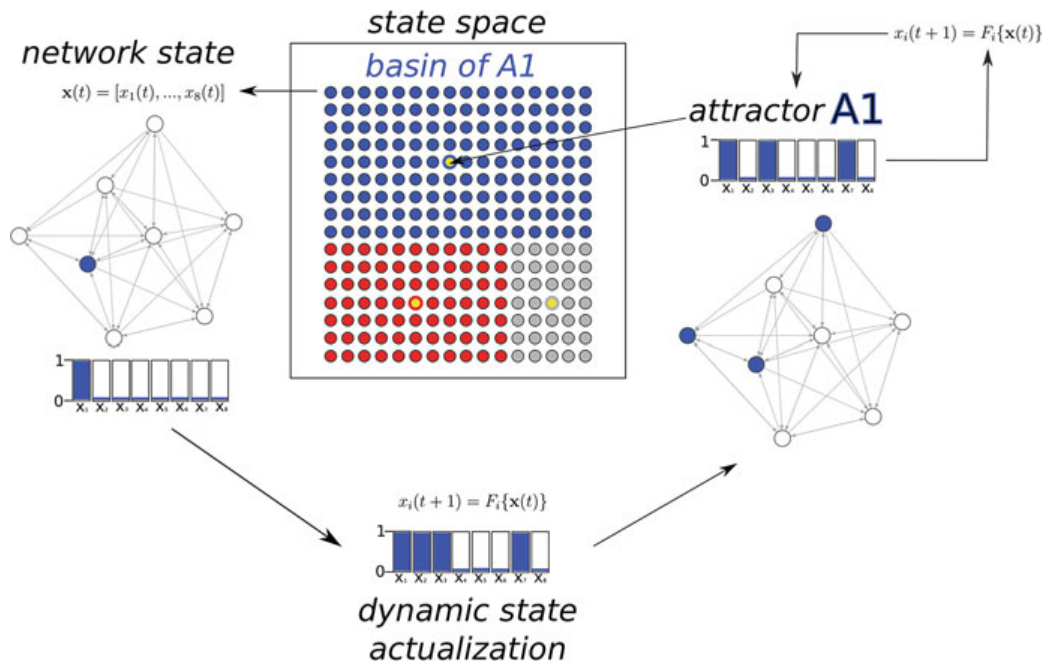
## 3 Methods

### 3.1 Definitions

We start by providing some definitions that are key to understanding how the conceptual framework of the epigenetic landscape is enclosed within the context of GRN dynamics.

A GRN is a directed graph composed of *nodes* and *edges*. Each node represents a gene (or another molecular type), and a state variable is assigned to each node to specify a *gene state*. Nodes are connected by edges that represent regulatory interactions among genes. In addition to the graph, which is simply represented as a wiring diagram (Fig. 3), the incoming regulatory interactions of each node are formalized as *updating rules* that determine the temporal activity of each node. In a Boolean GRN, gene activity changes are assumed in discrete time units. The gene state variables take only one of two values (0 when the gene is inactive, and 1 when the gene is active). The updating rules are logical propositions that determine the state of a given node in the next time step given the state of its incoming regulatory nodes (*see* Tables 3 and 5 for examples). The GRN is completely specified by the set of nodes and their corresponding updating rules [1]. Both the set of nodes and the updating rules are proposed based on experimental molecular data.

A gene configuration of the complete network (*network state*) is specified at each time step by the state of all the genes constituting the GRN. Given that gene states take only two values (0 or 1), a GRN with  $n$  genes can only take state values from a finite *state space* of  $2^n$  possible states. Starting from a given network state, the parallel implementation of the updating rules will determine a specific future network state. By the iterative application of the updating rules starting from a given network state it is possible for the network to reach either a stationary state that does not change after applying the updating rules (*fixed point attractor*). Alternatively, the networks can reach a close set of states from where the updating rules map to themselves (*cyclic attractor*). These stationary states are called *attractors*. If the updating rules are applied exhaustively starting from every possible network state, subsets of states converging to the same attractor can be determined. The regions in the state space converging to the same attractor are called *basins of attraction*. Importantly,



**Fig. 3** Schematic representation of an arbitrary GRN dynamical model. This diagram illustrates the general protocol of a dynamical analysis of a Boolean network. The state space contains all of the possible network states. Application of the logical functions to every network state eventually leads to an attractor. All the network states that lead to a particular attractor constitute its basin of attraction. The network state space is then structured by the basins of attraction of the different attractors

attractors represent activity configurations that are consistent with the regulatory logic specified in the updating functions—similar to the way gene activity profiles observed in vivo maintain specific cellular phenotypes. Thus, the attractor states of GRN models are expected to correspond to the cellular phenotypes observed in the real developmental systems being modeled. Interestingly, it has been shown in multiple developmental processes that this is indeed the case [14–16, 20].

As a result of the GRN dynamics, the state space is structured in a specific manner, which is reflected in the way the basins of attraction are organized. We refer to such structures as the *attractor landscapes*. The characterization of *attractor landscapes* is what allows us to explore the EL associated with the underlying GRN. The basins of attraction correspond to distinct differentiation states having different relative stability properties—akin to hierarchically organized valleys in the EL. Given the attractor landscape, the developmental pathways followed by a developmental system under normal or perturbed conditions can be explored by estimating the likelihood of specific single or sequential transitions among the basins of attraction (see below).

In Fig. 3 we schematically show the key concepts involved in modeling GRN dynamics. The implementation of the updating rules to a given network state will determine a specific network state for the next time step. Then, if one applies the updating functions

**Table 1**  
**General modeling concepts**

<b>Descriptive modeling</b>	
Model	A mathematical expression or computer algorithm that relates the values of one or more <i>responsive</i> (dependent) variables with the values of a set of <i>predictor</i> (independent) variables
Prediction	Calculated values of the responsive variables, which are assessed by taking specific values of the predictor variables as input to the model
Explanation	A predictor variable $x$ is said to <i>explain</i> 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 <i>explain</i> one variable in terms of the other—“ <i>correlation does not imply causation</i> ”
<b>Mechanistic modeling</b>	
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 <i>causal</i> processes considered in the model. The model is built by explicitly considering the processes that produce our observations

iteratively to this latter state, the network will eventually end up in an attractor state. In order to characterize the attractor landscape, we need to obtain all the attractors and their basins of attraction. After identifying the network’s attractors, we need to characterize their corresponding attraction basin. To further describe the attraction basins, we can retrieve the complete mapping structure linking every possible initial state to its corresponding attractor, which is known as the network’s transition table. Tables 1 and 2 summarize the concepts discussed in this section. In the following sections, we explain how to apply these concepts for systems biology modeling.

### **3.2 Dynamical Analysis of Boolean GRNs**

Before presenting the real case study of plant development, we first show how the dynamics of an arbitrary Boolean GRN can be

**Table 2**  
**GRN dynamical model concepts**

Concept	Definition
Node	Representation of a molecular species (gene, protein,...)
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

analyzed. Here, the modeler interested in a specific developmental system builds a network by choosing the set of genes and their regulatory interactions based on prior knowledge extracted from literature or directly from experiments (*see Note 3*). Interactions among nodes are encoded as logical propositions constituting the updating rules. For demonstrative purposes, here we will start by generating a random network as an example.

We follow a simplified protocol for constructing and analyzing a Boolean GRN model considering only the essential steps. For an in-depth modeling protocol that outlines how GRNs are proposed from scratch using experimental molecular data *see* Refs. **1** and **11**.

The present protocol includes only the next two essential steps:

1. Defining a set of nodes and their associated updating rules (*see Note 3*).
2. Characterizing the state space by identifying the attractors and their basins of attraction (*see Note 4*).

### 3.2.1 Building and Analyzing a Random Boolean GRN Model

1. *Defining the set of genes and updating functions:*  
Artificial random Boolean GRN models with specific properties can be easily generated using the software *BoolNet*. *BoolNet* includes the function `generateRandomNKNNetwork(n, k)`, which generates a random network with  $n$  genes, each with an updating rule having  $k$  input genes. Using this

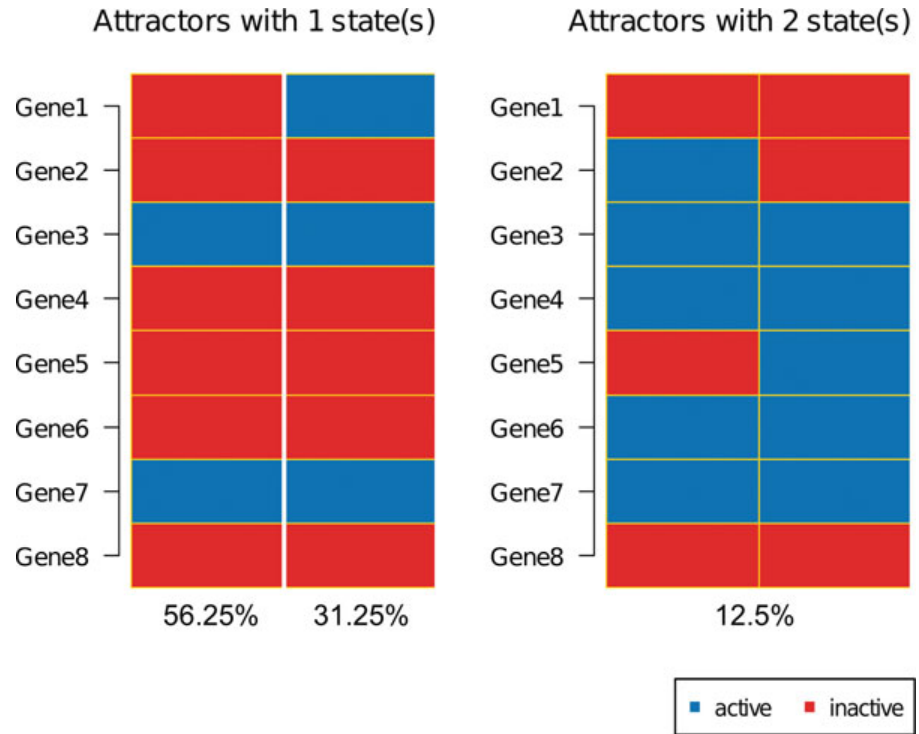
**Table 3**  
**Arbitrary Boolean GRN model**

List of state variables
$X = [\text{Gene1}, \text{Gene2}, \text{Gene3}, \text{Gene4}, \text{Gene5}, \text{Gene6}, \text{Gene7}, \text{Gene8}]$
Boolean functions
Gene1, $(\neg \text{Gene1} \ \& \ \text{Gene8}) \   \ (\text{Gene1} \ \& \ \neg \text{Gene8})$
Gene2, $(\neg \text{Gene5} \ \& \ \neg \text{Gene3}) \   \ (\text{Gene5} \ \& \ \text{Gene3})$
Gene3, $(\neg \text{Gene3}) \   \ (\neg \text{Gene8})$
Gene4, $(\neg \text{Gene1} \ \& \ \text{Gene4})$
Gene5, $(\text{Gene4} \ \& \ \neg \text{Gene5})$
Gene6, $(\text{Gene4}) \   \ (\text{Gene2})$
Gene7, $(\neg \text{Gene8}) \   \ (\neg \text{Gene7})$
Gene8, $(\text{Gene5} \ \& \ \text{Gene1})$

function, we created a network with eight genes, whose regulatory logic is determined by two input genes in each case. Note that the output network of this function will be different each time. The set of nodes and updating functions of the specific random network are included in Table 3. Alternatively, the updating rules can be loaded directly from a text file using the function `loadNetwork()`. To load a network in *BoolNet*, these rules have to be written using specific symbols: AND (&), OR (|), and NOT (!), in a syntax specified in the manual [25].

2. *Characterizing the state space: identifying the attractors and their basins of attraction*

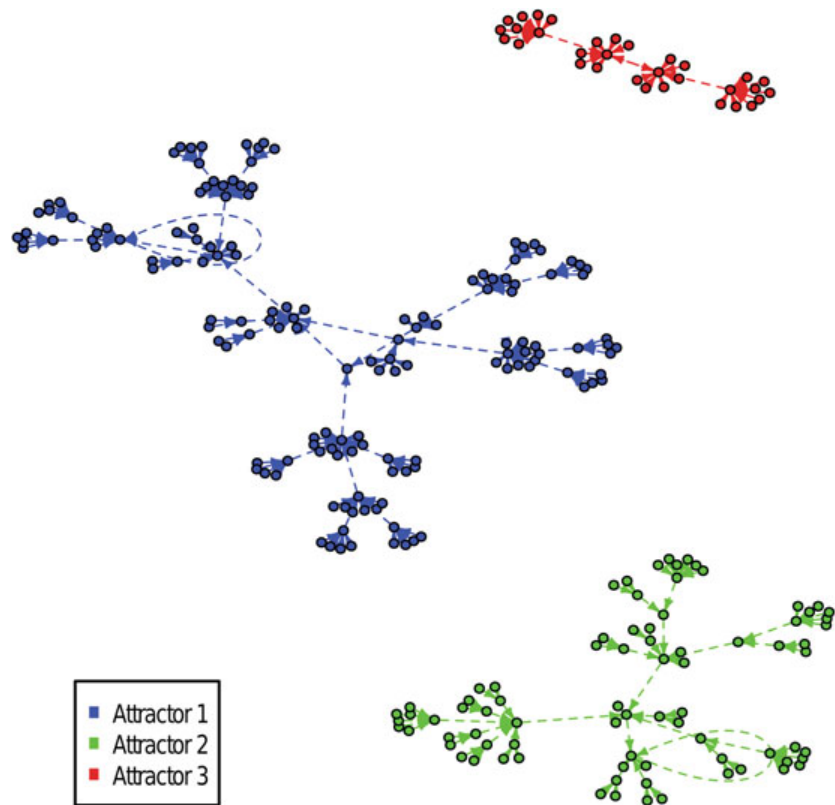
*BoolNet* package is implemented with functions that make the dynamical analysis of Boolean networks straightforward. Using *BoolNet*, the attractors of a given network can be easily obtained through the function `getAttractors(network)`. The default settings of this function recover the attractors of the given network by an exhaustive synchronous search. Intuitively, this means that it takes each of the network's  $2^n$  possible states as a starting point and follows their transitions until they reach an attractor. Specialized algorithms grant efficient exhaustive search in networks with up to 29 nodes. In bigger networks, heuristic search algorithms perform well by specifying the corresponding arguments to the same function [25, 26]. Once the attractors have been identified, a graphical depiction of how the state space is structured can be obtained with the function `plotAttractors()`.



**Fig. 4** Obtained attractors of the random GRN. The GRN recovered two fixed-point attractors and one cyclic attractor with two states. In the graph, blue color indicates expression or gene activation (1), while red color indicates no expression or inactivation (0)

This function outputs a graph as the one shown in Fig. 4, where rows correspond to genes and columns correspond to attractors. The activation state of each gene in the different attractors is color-coded (blue/red) cells represent active (inactive) genes). We can see that this network has an attractor landscape of three attractors, two of them are a fixed point and the third one is a cyclic attractor that oscillates between two states. The relative size of each attraction basin is shown under each attractor configuration (Fig. 4). This number represents the number of initial states that reach the given attractor divided by all the possible initial states.

To further describe the attraction basins, we recover the transition table, graphically represented as a state transition graph—i.e., a graph in which nodes are network states and arrows represent transition between nodes that result from the application of the updating rules. The *BoolNet* function `plotStateGraph()` generates a state transition graph, taking the computed attractors as input. Figure 5 shows the state transition graph for our example network. A different color is used for each attraction basin and attractors correspond to nodes having either only incoming edges (fixed-point attractor) or forming a loop (cyclic attractors with more than one state). The number of nodes in each basin of attraction defines its size.



**Fig. 5** State transition graph of the random GRN. Each node in this graph represents a network state connected to the state it takes after applying the updating rules. Each component of the graph corresponds to an attraction basin and is plotted in a different color

### 3.2.2 Building and Analyzing a ‘Real’ Boolean GRN Model from Experimental Data

We now turn to the *Arabidopsis* root SCN-GRN as an example of a dynamical analysis applied to a gene network based on experimental data [15]. As explained above, in order to model a developmental process as a Boolean GRN, the first step is to define the genes involved in the process and the regulatory interactions among them (Table 4). Azpeitia and collaborators proposed a Boolean network based on a thorough review of scientific literature (see Note 3). The obtained network recovers four attractors corresponding to the main cell types in the *Arabidopsis* root SCN [15].

It is important to recall that the four cell types of the root SCN can be distinguished by their expression pattern of the proposed set of genes (Fig. 2b). Defining the expected set of attractors is an indispensable step when building the GRN model because the attractors are used to validate the GRN (see Note 3). To further validate a proposed network model, it is important to simulate gene knockdown or mutations causing gene overexpression on the network and compare the results with experimental evidence (see Note 5). Azpeitia and collaborators followed this approach and showed that most predicted alterations to the stable configurations caused by mutant simulations are consistent with known empirical observations [15].



**Table 4**  
**Experimentally supported (real) interactions set for the root SCN-GRN, taken from [15]**

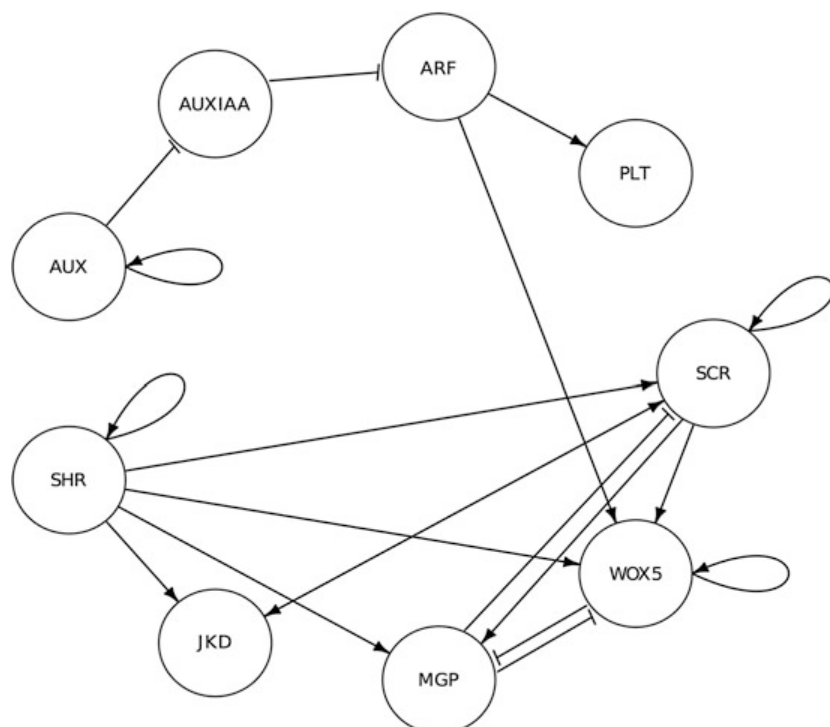
Interactions	Experimental evidence
SHR -> SCR	The expression of <i>SCR</i> is reduced in <i>shr</i> mutants. ChIP-QRTPCR experiments show that <i>SHR</i> directly binds in vivo to the regulatory sequences of <i>SCR</i> and positively regulates its transcription
SCR -> SCR	In the <i>scr</i> mutant background promoter activity of <i>SCR</i> is absent in the QC and CEI. A ChIP-PCR assay confirmed that <i>SCR</i> directly binds to its own promoter and directs its own expression
JKD -> SCR	<i>SCR</i> mRNA expression is lost in QC and CEI cells in <i>jdk</i> mutants from the early plantule stage onward
MGP- SCR	The double mutant <i>jdk/mgp</i> rescues the expression of <i>SCR</i> in QC and CEI cells, which is lost in the <i>jdk</i> single mutant
SHR -> MGP	The expression of <i>MGP</i> is severely reduced in the SHR background. <i>SHR</i> binding to <i>MGP</i> regulatory region has been confirmed by ChIP-PCR
SCR -> MGP	<i>SCR</i> directly binds to the <i>MGP</i> promoter, and <i>MGP</i> expression is reduced in the <i>scr</i> mutant background
SHR -> JKD	The post-embryonic expression of <i>JKD</i> is reduced in <i>shr</i> mutant roots
SCR -> JKD	The post-embryonic expression of <i>JKD</i> is reduced in <i>scr</i> mutant roots
SCR -> WOX5	<i>WOX5</i> is not expressed in <i>scr</i> mutants
SHR -> WOX5	<i>WOX5</i> expression is reduced in <i>shr</i> mutants
ARF(MP)-> WOX5	<i>WOX5</i> expression is rarely detected in <i>arf(mp)</i> or <i>arf(bdl)</i> mutants
ARF-> PLT	<i>PLT1</i> mRNA is overexpressed under ectopic auxin addition. <i>PLT1</i> & <i>2</i> mRNAs are absent in the majority of <i>arf(mp)</i> embryos
Aux/IAA- ARF	Overexpression of <i>Aux/IAA</i> genes represses the expression of <i>ARF(DR5)</i> both in the presence and absence of auxin. Domains III & IV of <i>Aux/IAA</i> genes interact with domains III & IV of <i>ARF</i> stabilizing the dimerization that represses <i>ARF</i> transcriptional activity
Auxin-  aux/IAA	Auxin application destabilizes <i>Aux/IAA</i> proteins. <i>Aux/IAA</i> proteins are targets of ubiquitin-mediated auxin-dependent degradation

### 3.2.3 Dynamical Analysis with BoolNet

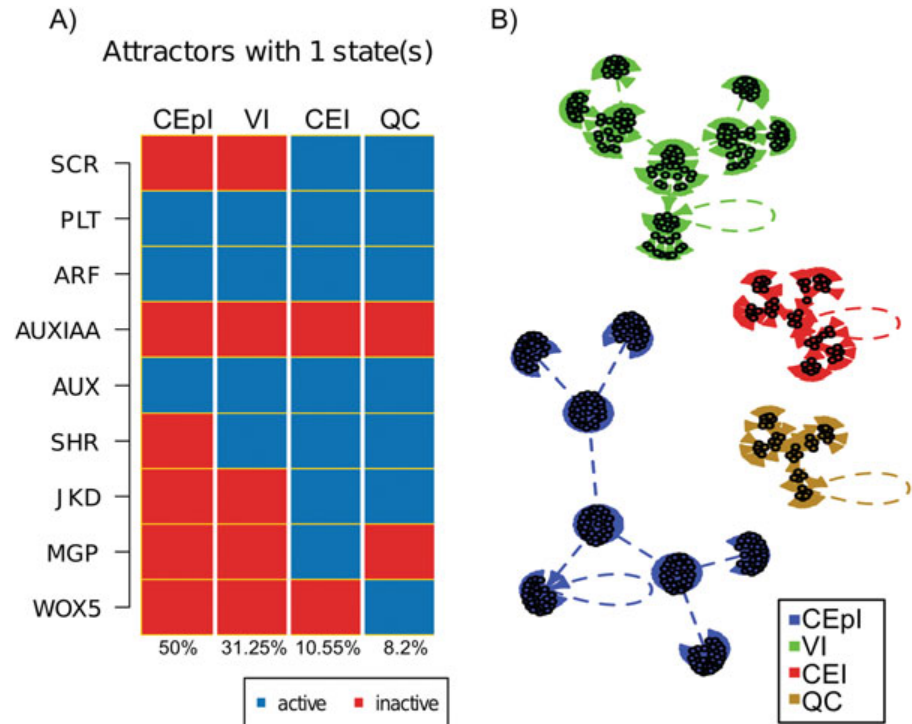
We use the logical rules proposed by Azpeitia and collaborators (summarized in Table 5) to perform the analysis in *BoolNet*. The rules are loaded by simply using the function `loadNetwork("root_SCN.txt")`, where `root_SCN.txt` is a text file containing the logical rules, which should be saved in the current R working directory. Figure 6 shows the wiring diagram of the root SCN-GRN, a graphical representation of the Boolean rules used to define the network.

**Table 5**  
***Arabidopsis* root SCN-GRN Boolean functions**

List of state variables
$X = [\text{SCR}, \text{PLT}, \text{ARF}, \text{AUXIAA}, \text{AUX}, \text{SHR}, \text{JKD}, \text{MGP}, \text{WOX5}]$
Boolean functions
$\text{SCR}, (\text{SHR} \ \& \ \text{SCR} \ \& \ !\text{JKD} \ \& \ !\text{MGP}) \   \ (\text{SHR} \ \& \ \text{SCR} \ \& \ \text{JKD} \ \& \ \text{MGP}) \   \ (\text{SHR} \ \& \ \text{SCR} \ \& \ \text{JKD} \ \& \ !\text{MGP})$
PLT, ARF
ARF, !AUXIAA
AUXIAA, !AUX
AUX, $\text{AUX} \   \ !\text{AUX}$
SHR, SHR
JKD, $\text{SHR} \ \& \ \text{SCR}$
MGP, $\text{SHR} \ \& \ \text{SCR} \ \& \ !\text{WOX5}$
$\text{WOX5}, (\text{ARF} \ \& \ \text{SHR} \ \& \ \text{SCR} \ \& \ !\text{MGP} \ \& \ !\text{WOX5}) \   \ (\text{ARF} \ \& \ \text{SHR} \ \& \ \text{SCR} \ \& \ !\text{MGP} \ \& \ \text{WOX5}) \   \ (\text{ARF} \ \& \ \text{SHR} \ \& \ \text{SCR} \ \& \ \text{MGP} \ \& \ \text{WOX5})$



**Fig. 6** Wiring diagram of the *Arabidopsis* root SCN-GRN. Single-cell root SCN GRNs proposed in [15]. Nodes represent genes and edges represent regulatory interactions among them. The effect of the interaction is symbolized with arrows and flat arrows for activating or repressing interactions, respectively. Note that when using *BoolNet*'s function `plotNetworkWiring` the resulting plot does not distinguish between activating and repressing interactions



**Fig. 7** Attractors and state transition graph of the root SCN-GRN. (a) Genetic configuration of the four attractors obtained from the root SCN-GRN. These attractors correspond to the experimental gene expression profiles of the cell/types found at the root SCN. (b) State transition graph, representing each attractor's basin of attraction

Once the network is loaded into *BoolNet*, we follow the same dynamical analysis presented above for the random network. In summary, we find the network's attractors and describe their corresponding attraction basins. The output of the functions `getAttractors()` and `plotAttractors()` indicate that the network converges to four attractors (Fig. 7a). The recovered attractors correspond to the expression profiles of the four cell types found in the *Arabidopsis* root SCN (Table 6): VI, CEI, QC and CEpl [15]. This suggests that cell-type gene-expression patterns in the root SCN result from the restrictions imposed by the uncovered GRN developmental module. As explained above, each attractor's basin of attraction can be graphically represented using the function `plotStateGraph()` (Fig. 7b).

### 3.3 Modeling the EL

In a systems biology conceptual framework, a cell is represented as a dynamical system governed by an underlying GRN. The trajectories leading to the attractors can be naturally associated to the valleys depicted in the EL metaphor proposed by Waddington [27]. Although the idea underlying the EL metaphor is intuitively easy to understand, it cannot be directly operationalized with the conventional Boolean GRN formalism. GRN attractors are steady states, this implies that once a network reaches an attractor state it

**Table 6**  
**Gene expression profiles of the root SCN cell types (expected attractors)**

Cell type	PLT	Auxin	ARF	Aux/IAA	SHR	SCR	JKD	MGP	WOX5
QC	1	1	1	0	1	1	1	0	1
VI	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

will stay there indefinitely unless an external force moves the system out of such a state. In real developmental processes, however, cells transit from one cell state (attractor) to another. A cell can change its state by two nonexclusive mechanisms. On the one hand, cell state changes can be driven by intrinsic stochastic fluctuations of the molecular system. On the other hand, cells can change state in response to extrinsic signaling factors [28, 29]. Both mechanisms can be modeled within GRN dynamics. Stochastic noise can be introduced to the model causing cells to jump around the landscape without requiring any parameter changes. The effect of extrinsic factors can be modeled by altering the parameters of the network structure and thus changing the attractors landscape [30]. Thus, in order to characterize the EL associated with a Boolean GRN, it is necessary to extend the discrete network model in order to explore transitions between attractors.

Here we consider a modeling extension to include stochastic intrinsic perturbations. This is achieved by randomly perturbing the state of the genes in the network causing “jumps” among attractors. This scenario assumes that developmental transitions are the natural consequence of the regulatory restrictions themselves and not of the signaling mechanisms. The fact that genetic regulatory interactions represented in a GRN are biochemical reactions subject to stochastic fluctuations makes the inclusion of stochasticity to any proposed GRN model a valid assumption. Given the nonlinear restrictions imposed by the underlying regulatory network, the potential jumping patterns among network states will not be equally likely, in spite of unbiased stochasticity. This interacting effect between nonlinearity and stochasticity enables the discovery of nontrivial, robust patterns of transitions.

Based on the considerations above, we proceed by introducing a general strategy for the practical extension of a validated Boolean GRN in order to produce an EL model. The framework is exemplified in the next section, and it comprises three steps:

1. Computational simulation of cell state changes in response to perturbations generated by introducing stochasticity into the Boolean dynamics.
2. Analysis of the prevailing paths of cell fate change for estimating an interattractor transition probability matrix.
3. Characterization of the temporal evolution of the probability distribution over attractor states.

### 3.3.1 *Introducing Stochasticity to the Boolean Dynamics*

Stochastic noise can be introduced to a GRN in different ways (*see Note 6*). Here we use the model known as stochasticity in nodes (SIN) [31], which introduces stochasticity by considering a fixed probability for a gene to disobey its updating rule. In other words, under the SIN model, even if the updating rules of a gene  $X$  imply that it should be active (or inactive) in the next time step, there is a certain probability of the contrary output to occur. Under this stochastic dynamics, a given initial configuration will no longer converge to the same attractor every time. The probability of the network to pass from one network state to another can be estimated by iterating the stochastic rule a large number of times (at least 1000 times) and estimating the frequency of the interstate transitions. The estimated transition probabilities can then be used to study the behavior of the system and to make statistical predictions of cell state transitions.

### 3.3.2 *Building the Interattractor Transition Matrix*

The state transition probabilities we are interested in are those that occur between network states belonging to different basins of attraction. The probabilities of passing from any attractor to any another are arranged into an interattractor transition matrix (IATM). In order to estimate the IATM, we first introduce stochasticity to the GRN and iterate the stochastic functions for every network state a large number of times. For every iteration, we must store the basin of attraction the original state belonged to, and the basin of attraction it reaches after introducing stochasticity and applying the updating rules. Using this simulated state transition, the interattractor transition probabilities are calculated as the relative number of times the network “jumps” to every possible attraction basin starting from the one the initial state belongs to.

### 3.3.3 *Downstream Analyses to Characterize the EL*

Once the IATM has been computed, downstream analyses can be performed in order to uncover the underlying EL structure emerging from the regulatory restrictions. Some of these analyses include computing the temporal order of attractor attainment and the attractor relative stability and global ordering. We will explain the basic ideas underlying these analyses.

### 3.3.4 *Temporal Sequence of Attractor Attainment*

Computing the temporal sequence of attractor attainment is a basic approach to developmental phenomena that involves uncovering

the regulatory basis for the typical temporal sequence of cell-type acquisition. To perform it, it is necessary to have a hypothesis about the initial distribution of cell-types. This means that a given process begins with a population of cells distributed across the available cell-types. Further throughout the process, some of those cell types differentiate into other cell types. In the present model, the changes between cell-types are encoded in the IATM, and the process of cell population differentiation is simulated by multiplying an attractor distribution vector times the IATM. In order to perform this analysis, an initial cell-type (attractor) distribution vector must be defined.

This is a vector  $P_X(t_0) = (p_1(t_0), p_2(t_0), \dots, p_K(t_0))$ , where  $p_i(t_0)$  represents the probability of the network being in attractor  $i$  at the initial time  $t = 0$ . From a cell population point of view, the probability of attractor  $i$  at any time is interpreted as the percentage of cell type  $i$  in the population at that time. Having the initial distribution vector, the dynamics of each attractor's probability in time is simulated by iteratively multiplying the distribution vector times the IATM a certain number of time steps. As proposed in [17], the succession of attractors' probability *maxima* then corresponds to an intrinsic explanation for the emerging temporal order observed during a developmental process.

### 3.3.5 Relative Attractor Stability and Global Ordering

During development, the zygote differentiates into different cell types in a defined and robust order, which makes most steps in the cell differentiation process irreversible. GRNs, proposed as a representation of the genetic mechanism underlying cell differentiation, are expected to recover the observed sequence of attractors in the presence of noise. Zhou and collaborators proposed a method to calculate the global attractor ordering of a given GRN based on an attractor's relative stability [19]. Relative stability of attractors reflects the relative ease for transitioning from an attractor (A) to another attractor (B) given a certain degree of stochastic noise, i.e., the probability of passing from attractor A to attractor B. The relative stabilities in a GRN are expected to be asymmetric between any pair of attractors, which gives the attractor ordering directionality (i.e., irreversibility during cell differentiation).

Relative stabilities of a GRN attractor can be calculated by the mean first passage time (MFPT) from a given IATM as proposed by Zhou and collaborators [17]. The MFPT between attractors A and B is the expected number of time steps until reaching attractor B starting from attractor A. MFPTs can be calculated either by implementing the matrix-based algorithm proposed in [32] or by means of numerical simulation. After defining the MFPT among every attractor pair, a net transition rate ( $d_{i,j}$ ) between attractor  $i$  and  $j$  is defined in terms of the MFPT as follows:

$$d_{i,j} = \frac{1}{\text{MFPT}_{i,j}} - \frac{1}{\text{MFPT}_{j,i}}$$

The attractor global ordering can be obtained by calculating the transition rate among all the attractors. The consistent global ordering of the attractors is given by the attractor permutation in which all transitory net transition rates from an initial attractor to a final attractor are positive, as proposed in [19]. This can be illustrated by constructing a network using the transition rates matrix as an adjacency matrix and highlighting the positive transition rates. In the global ordering network nodes are the attractors and the arrows connecting them are the transition rates among them. From such a network one can recover the global ordering looking for the path that connects all the attractors through positive transition rates.

We have implemented all the modeling extensions introduced in this section so that they can be applied directly to the output of the dynamical analyses presented in the previous sections. In what follows we exemplify their use.

### 3.3.6 Implementing the EL Protocol

We have coded functions in R for a practical implementation of the complete framework of EL modeling, applicable to a dynamically analyzed Boolean GRN (see Subheading 3.2). These functions recover the steps explained above for EL modeling, these are: calculation of the IATM, temporal sequence of attractor attainment, and attractors global ordering. We apply these functions to the *Arabidopsis* root SCN-GRN analyzed above, and interpret the results.

### 3.3.7 Calculating the IATM

The first step is the calculation of the IATM. For this purpose we coded the function `Implicit.InterAttractor.Simulation` (`Network`, `P.error`, `Nreps`), which estimates the IATM by simulating stochastic node perturbations in the network dynamics. It takes as inputs a Boolean GRN, an error probability (see Note 7) and the desired number of iterations (1000 or more) to be performed over each network state. The function works as follows:

1. Recovers the state space of the network and the attraction basin each state belongs to.
2. Computes the state transition for every initial state.
3. Changes the state of every gene considering the given error probability.
4. Identifies the attraction basin of the resultant perturbed transitory state.
5. Records every change of attraction basin.

The steps are repeated the indicated number of times and the frequency of observed transitions among all the attractor basins is estimated (for details, see [18]). Using this function, we estimate the IATM for the root SCN-GRN considering an error probability

**Table 7**  
**Interattraction probability matrix for the root SCN-GRN, with an error probability of 0.05**

	CepI	VI	CEI	QC
CepI	0.9498	0.0479	0.0011	0.0012
VI	0.0513	0.9032	0.0239	0.0216
CEI	0.0509	0.0745	0.8141	0.0605
QC	0.0502	0.0608	0.0796	0.8094

of 0.05 and 1000 repetitions. The resulting IATM is shown in Table 7. Note that due to stochasticity interattractor probabilities in different simulations will vary slightly.

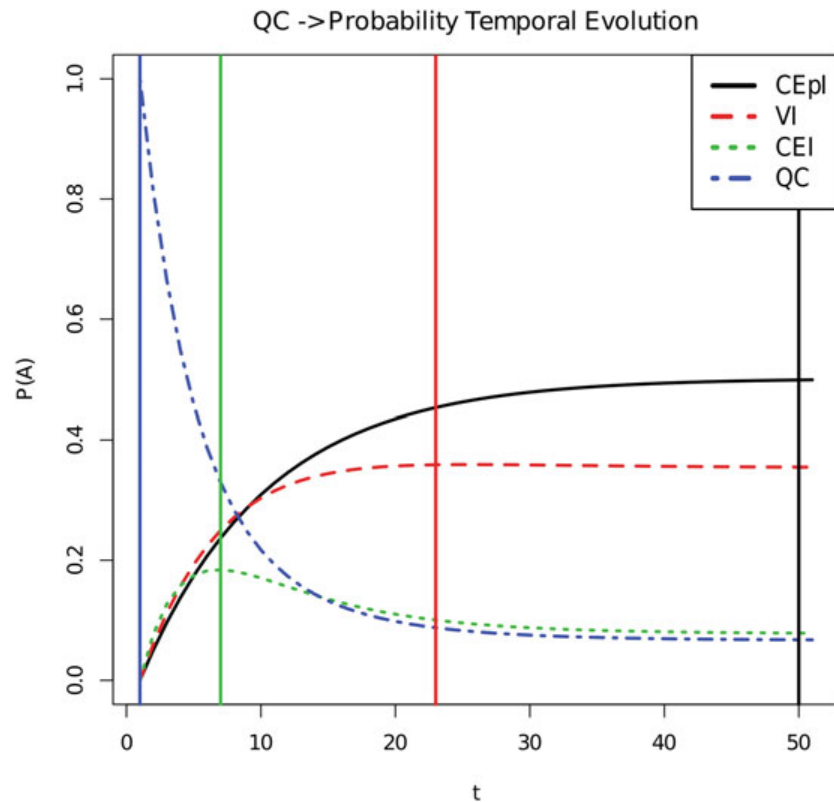
### 3.3.8 Estimating the Attractor Temporal Evolution

Once having calculated the IATM, the temporal sequence of attractor attainment can be easily estimated. For a practical implementation, we coded the function.

`Plot.Probability.Evolution(TPM, Initial, AttrsNames, timeF)`. It takes as inputs an IATM, the name of the initial attractor, the names of all the network's attractors in the same order as they are obtained by `getAttractors`, and the number of times to iterate the process. As mentioned above, we need to provide an initial attractor distribution to calculate the temporal evolution. In this function, the initial attractor provided by the user is assumed to be the only cell type present in the initial distribution. The result of the function is a matrix with the attractor's probability distribution corresponding to each time-step, and a plot that illustrates the dynamics of each attractor's probability and their temporal attainment order.

We calculated the temporal order of cell-type attainment in the root SCN, using as input the IATM calculated above with an initial distribution of quiescent center cells for 50 time-steps: `Plot.Probability.Evolution(IAT_5, Initial="QC", AttrsNames=c("CEpI", "VI", "CEI", "QC"), timeF=50)`. In the resulting plot (Fig. 8) the obtained temporal order of cell types follows sequentially: QC, CEI, VI, and CEpI. The biological interpretation of this result is not straightforward since the root SCN has a continuous production of cells and the characteristic cell-types follow a topological order rather than a temporal one. Alvarez-Buylla and collaborators proposed one of the first methodological frameworks for the exploration of the EL. They proposed a GRN underlying the cell-types of early flowering in *A. thaliana*, and through an IAT approach they recovered the temporal sequence of cell types observed during flower development [17].





**Fig. 8** Temporal sequence of cell–fate attainment pattern in the root SCN-GRN, starting with quiescent center cells. Each line in the plot corresponds to each attractor's probability of occurrence through time, the vertical lines indicate the time of the maximum probability for the corresponding color-coded attractor

The last analysis we present here is the global ordering of attractors, for which it is necessary to calculate the MFPT and transition rate matrices. We have created the functions `Calculate.MFPT.Matrix` and `MFPT.Transition.Rates`, which calculate the MFPT matrix among attractors of a BN and their corresponding transition rates, respectively. The former function calculates the MFPT taking as inputs an IATM and the names of the obtained attractors. The latter function takes as input the calculated MFPT. Once we have calculated the inter attractor transition rates, we can obtain the attractor global ordering. We created the function `Plot.Attractor.Global.Ordering`, which takes the transition rates matrix as input and creates a network from the attractors' transition rates that highlights the positive transition rates as red arrows. As mentioned above, the attractors global ordering is the path in the network that passes through every attractor by positive transition rates.

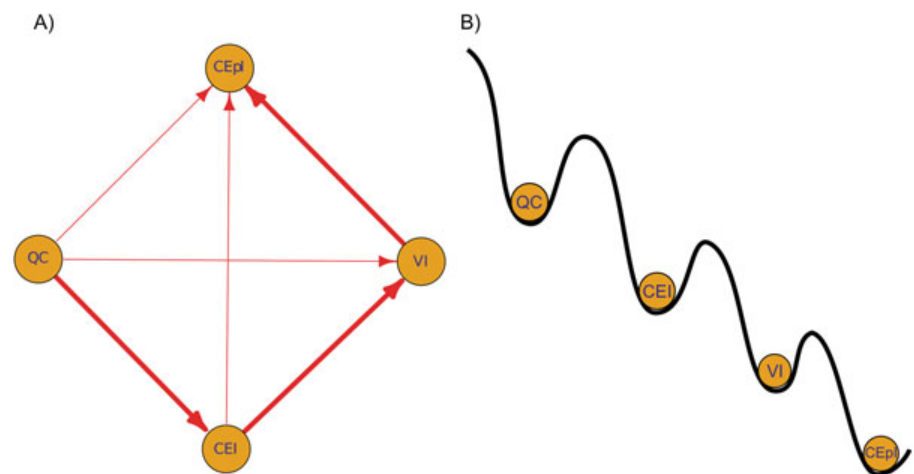
Using these functions and the IATM calculated above, we calculated the MFPT matrix (Table 8) and the transition rates matrix (Table 9) of the root SCN-GRN. Finally, we plotted the attractors' global ordering network (Fig. 9a) using the latter matrix as input. In Fig. 9, we can see that the temporal flow of the root

**Table 8**  
**Mean first passage time among attractors of the root SCN-GRN**

	<b>CEpI</b>	<b>VI</b>	<b>CEI</b>	<b>QC</b>
CEpI	0	20.52	80.14	90.49
VI	20.38	0	63.26	72.76
CEI	20.40	16.34	0	59.53
QC	20.45	17.41	47.48	0

**Table 9**  
**Transition rates among attractors of the root SCN-GRN**

	<b>CEpI</b>	<b>VI</b>	<b>CEI</b>	<b>QC</b>
CEpI	0	-0.0027	-0.0387	-0.0406
VI	0.0027	0	-0.0455	-0.0432
CEI	0.0387	0.0455	0	-0.0040
QC	0.0406	0.0432	0.0040	0



**Fig. 9** Global order of the root SCN-GRN attractors. **(a)** Global order network of the SCN attractors. The global order corresponds to the path of positive interactions passing through all of the attractors, the corresponding path is highlighted in the image. **(b)** Graphical interpretation of the landscape inferred from the attractor’s global order

SCN attractors follows the order: QC -> CEI -> VI-> CepI, which corresponds to the probability evolution of attractor attainment starting from QC. From the attractor global ordering, we can infer a representation of the EL as rifts and valleys as shown in Fig. 9b.

The transitions uncovered by the EL extension indicate that progenitor cells in the root stem cell niche are prone to differentiate into specific cell type given their current state, and as a natural consequence of the regulatory constraints. For example, our results would suggest that cells from the QC have a natural tendency to transit to the CEI state. Such prediction could be an interesting starting point to propose hypotheses for the analysis and interpretation of, for example, single-cell expression data, which is becoming more available [33, 34]. The predictions can also be relevant for in vitro differentiation experiments, given that the current model is not considering any additional, tissue-level restriction to the intrinsic regulatory interactions. Ultimately, the modeling protocols presented here should ideally be integrated with experiments. Interdisciplinary systems biology approaches considering the interplay among model building, analysis/simulation, and experimentation enable the discovery and interpretations of interesting and counterintuitive observed behaviors (see, for example 20).

---

## 4 Conclusion and Outlook

Computational modeling is a useful approach for understanding biological developmental processes. Integration of molecular genetic and genomic information in GRNs, allows the postulation of computer models that explore the epigenetic landscape, as theoretically proposed by C.H. Waddington [5]. Cross talk between experimental, theoretical, and computational approaches to biological systems is necessary for an integral comprehension of developmental processes.

In this chapter we present a framework for the analysis of the epigenetic landscape associated with a GRN. The proposed methodology is based on the Boolean modeling of GRNs, a useful scheme for explaining the different gene expression patterns among cell types as attractors of an underlying GRN. Moving forward from the Boolean dynamical analysis, the epigenetic landscape associated with a GRN is characterized by introducing stochasticity to the model and by measuring the probability of transitioning among attractors. The introduction of stochasticity makes possible the exploration of the epigenetic landscape implicit in the proposed GRN. We include a comprehensive exposition of our proposed methodology, together with its computational implementation. We expect the toolkit and conceptual interpretation put forward here to be useful resources for the systems biology community interested in modeling plant developmental processes.

## 5 Notes

1. A node in a GRN represents a variable describing the system behavior. In a general dynamical model variables are chosen both for practical and fundamental considerations. On the one hand, it is preferable to choose a variable whose value is easy to approximate experimentally. On the other hand, the modeler is commonly interested in finding variables that reflect relevant characteristics about the functional behavior of the system, and which are involved in the mechanistic basis underlying the latter. Although in a GRN, nodes are generically referred to as genes, these can represent any cellular element whose activity has a strong influence on the cellular phenotype. These elements commonly are proteins or protein complexes, signaling molecules, miRNAs, or groups of various elements representing a process. Due to the generality of the underlying modeling apparatus (i.e., a dynamical system), any kind of variable that changes with time can be considered a node in the network.
2. For a given GRN, an attractor is a network state which, if taken as an input for the Boolean functions, either does not move the network to another state (steady state attractor), or only moves through the same set of states (cyclic attractor). In other words, it is a network state (or sequence of states) that does not change in time. Every network state not belonging to an attractor is a transitory state that eventually leads to an attractor.
3. Boolean GRN model construction is an intuitive process in which experimental regulatory interactions are transferred into Boolean syntax. This process involves the integration of large amounts of experimental evidence obtained from the literature to be formalized as Boolean rules or truth tables. Experimental natural-language expressions can be stated as Boolean functions in a straightforward manner, as shown in Table 4. We will use an example from the root SCN-GRN to illustrate this process:

Roots with *shr* or *scr* knockdown show reduced expression levels of JKD. This suggests that: SHR and SCR are positive regulators of JKD.

The latter statement can be transformed easily into a Boolean function such as:

$$\text{JKD} = \text{SCR AND SHR}$$
4. The dynamical analysis of a Boolean network recovers the network's steady states (attractors). In GRN models of developmental modules, attractors are considered as an abstract

representation of cellular phenotypes, experimentally accessible through gene expression profiling. In order to validate a proposed GRN, the expected attractors have to be defined, i.e., the expression profiles of the cell-types of interest must be identified from experimental evidence and coded into a Boolean vector. For a GRN to be experimentally validated, the attractors it recovers must match the expected attractors. Nevertheless, it is important to keep in mind that different networks converge to the same set of attractors. Experimental evidence should weight the evaluation process of competing models.

5. The BoolNet package has straightforward ways to implement knockout and overexpression simulation experiments. Specifically, the genes within the network can be set to a fixed value (0 for knockout, and 1 for overexpression). Calculations can then be performed on the modified network, with the only difference that the assigned value, and not the one generated by corresponding transition function, will be used through the simulation. 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 dynamical analysis, such as attractor identification, can be performed on this new-perturbed network.
6. When working with Boolean GRNs, stochasticity can be introduced either by the SIN method (*see* Subheading 3.3.1) or by the stochasticity in function (SIF) method. In the SIF method, stochasticity is modeled at the level of biological functions (i.e., Boolean functions in the GRN), i.e., implicitly behaving contrary to what the Boolean function indicates and not just flipping the state of a gene as in the SIN model (for details *see* refs. 26, 31).
7. The level of noise (error probability) used in a stochastic model determines the behavior that will be recovered. When introducing stochasticity to a Boolean network, very small levels of noise are not strong enough to make the system leave an attractor so no state transitions will be recovered. On the other hand, high levels of noise cause the system to jump among attractors completely randomly losing the information contained in the network. An appropriate noise level shows a nontrivial behavior in which there are state changes following the logic of the network. Levels of error probability used for Boolean GRNs range normally from 0.01 to 0.1 [16, 17], but different values should be tested.

## References

- Davila-Velderrain J, Martinez-Garcia JC, Alvarez-Buylla ER (2015) Descriptive vs. mechanistic network models in plant development in the post-genomic era. *Methods Mol Biol* 1284:455–479
- Álvarez-Buylla ER, Dávila-Velderrain J, Martínez-García JC (2016) Systems biology approaches to development beyond bioinformatics: nonlinear mechanistic models using plant systems. *Bioscience* 66(5):371–383
- Forgacs G, Newman SA (2005) *Biological physics of the developing embryo*. Cambridge University Press, Cambridge
- 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* 21:108–117
- Waddington CH (1957) *The strategy of genes*. George Allen & Unwin, Ltd., London
- Davila-Velderrain J, Martinez-Garcia JC, Alvarez-Buylla ER (2015) Modeling the epigenetic attractors landscape: toward a post-genomic mechanistic understanding of development. *Front Genet* 6:160
- Alvarez-Buylla ER, Balleza E, Benítez M et al (2008) Gene regulatory network models: a dynamic and integrative approach to development. *SEB Exp Biol Ser* 61:113
- Kauffman SA (1969) Metabolic stability and epigenesis in randomly constructed genetic nets. *J Theor Biol* 22(3):437–467
- Alvarez-Buylla ER, Benítez M, Davila EB et al (2007) Gene regulatory network models for plant development. *Curr Opin Plant Biol* 10(1):83–91
- Davila-Velderrain J, Martinez-Garcia JC, Alvarez-Buylla ER (2016) Dynamic network modelling to understand flowering transition and floral patterning. *J Exp Bot* 67(9):2565–2572
- Azpeitia E, Davila-Velderrain J, Villarreal C, Alvarez-Buylla ER (2014) Gene regulatory network models for floral organ determination. *Methods Mol Biol* 1110:441
- Kaplan D, Glass L (2012) *Understanding nonlinear dynamics*. Springer, New York
- Glass L, Kauffman SA (1973) The logical analysis of continuous, non-linear biochemical control networks. *J Theor Biol* 39(1):103–129
- 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(11):2923–2939
- 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(1):1
- Benítez M, Espinosa-Soto C, Padilla-Longoria P, Alvarez-Buylla ER (2008) Interlinked nonlinear subnetworks underlie the formation of robust cellular patterns in *Arabidopsis* epidermis: a dynamic spatial model. *BMC Syst Biol* 2(1):1
- Álvarez-Buylla ER, Chaos Á, Aldana M et al (2008) Floral morphogenesis: stochastic explorations of a gene network epigenetic landscape. *PLoS One* 3(11):e3626
- Davila-Velderrain J, Juarez-Ramiro L, Martinez-Garcia JC, Alvarez-Buylla ER (2015) Methods for characterizing the epigenetic attractors landscape associated with Boolean gene regulatory networks. *arXiv preprint arXiv:1510.04230*
- Zhou JX, Samal A, d'Hérouël AF, Price ND, Huang S (2016) Relative stability of network states in Boolean network models of gene regulation in development. *Biosystems* 142:15–24
- Pérez-Ruiz RV, García-Ponce B, Marsch-Martínez N et al (2015) XAANTAL2 (AGL14) is an important component of the complex gene regulatory network that underlies *arabidopsis* shoot apical meristem transitions. *Mol Plant* 8(5):796–813
- Cui H, Levesque MP, Vernoux T, Jung JW et al (2007) An evolutionarily conserved mechanism delimiting SHR movement defines a single layer of endodermis in plants. *Science* 316:421–425
- Levesque MP, Vernoux T, Busch W, Cui H et al (2006) Whole-genome analysis of the SHORT-ROOT developmental pathway in *Arabidops*. *PLoS Biol* 4:e143
- Sarkar AK, Luijten M, Miyashima S, Lenhard M, Hashimoto T, Nakajima K et al (2007) Conserved factors regulate signalling in *Arabidopsis thaliana* shoot and root stem cell organizers. *Nature* 446:811–814
- Stahl Y, Wink RH, Ingram GC, Simon R (2009) A signaling module controlling the stem cell niche in *Arabidopsis* root meristems. *Curr Biol* 19:909–914

25. Müssel C, Hopfensitz M, Kestler HA (2010) BoolNet—an R package for generation, reconstruction and analysis of Boolean networks. *Bioinformatics* 26(10):1378–1380
26. 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
27. Bhattacharya S, Zhang Q, Andersen ME (2011) A deterministic map of Waddington's epigenetic landscape for cell fate specification. *BMC Syst Biol* 5:85
28. Moris N, Pina C, Arias AM (2016) Transition states and cell fate decisions in epigenetic landscapes. *Nat Rev Genet* 17(11):693–703
29. Martinez-Sanchez ME, Mendoza L, Villarreal C, Álvarez-Buylla ER (2015) A minimal regulatory network of extrinsic and intrinsic factors recovers observed patterns of CD4+ T cell differentiation and plasticity. *PLoS Comput Biol* 11:e1004324
30. Davila-Velderrain J, Villarreal C, Alvarez-Buylla ER (2015) Reshaping the epigenetic landscape during early flower development: induction of attractor transitions by relative differences in gene decay rates. *BMC Syst Biol* 9(1):20
31. Garg A, Mohanram K, Di Cara A, De Micheli G, Xenarios I (2009) Modeling stochasticity and robustness in gene regulatory networks. *Bioinformatics* 25(12):i101–i109
32. Sheskin TJ (1995) Computing mean first passage times for a Markov chain. *Int J Math Educ Sci Technol* 26(5):729–735
33. Brennecke P, Anders S, Kim JK, Kołodziejczyk AA, Zhang X, Proserpio V, Baying B, Benes V, Teichmann SA, Marioni JC, Heisler MG (2013) Accounting for technical noise in single-cell RNA-seq experiments. *Nat Methods* 10(11):1093–1095. <https://doi.org/10.1038/nmeth.2645>
34. Efroni I, Ip P-L, Nawy T, Mello A, Birnbaum KD (2015) Quantification of cell identity from single-cell gene expression profiles. *Genome Biol* 16(1):9. <https://doi.org/10.1186/s13059-015-0580-x>

# B Cancer in attractor landscape modeling: a systems biology perspective of the disease

Capítulo publicado como parte del libro *Cancer a complex disease*, editado por Octavio Miramontes y Elena Álvarez-Buylla Roces, publicado en 2018 por CopIt'arXives.



## CANCER IN ATTRACTOR LANDSCAPE MODELING: A SYSTEMS BIOLOGY PERSPECTIVE OF THE DISEASE

*Jose Luis Caldu-Primo<sup>†c</sup>, Jose Davila-Velderrain<sup>±</sup>, Juan Carlos Martinez-Garcia<sup>‡</sup> and  
Elena R. Alvarez-Buylla<sup>†c\*</sup>*

<sup>†</sup>Instituto de Ecología, UNAM, Mexico City, Mexico

<sup>‡</sup>Departamento de Control Automatico, CINVESTAV-IPN, Mexico City, Mexico

<sup>c</sup>Centro de Ciencias de la Complejidad C3, UNAM, Mexico City, Mexico

<sup>±</sup>MIT Computer Science and Artificial Intelligence Laboratory, Cambridge, Massachusetts, USA

Cancer is a degenerative chronic disease that can be interpreted as a robust process intrinsic to human development. Traditional cancer research has considered it a genetic disease and focused on finding genetic mutations causing it. This genocentric approach has inherent limitations as it does not take into account the complex processes involved in the determination of phenotypes from a given genotype. In the field of systems biology, it has been established that cell lineage commitment and differentiation are governed by the dynamics of an underlying complex gene regulatory network (GRN). In this way, development and cellular differentiation can be understood using the epigenetic attractors landscape metaphor as originally proposed by C. H. Waddington. From this perspective it is possible to study the mechanisms underlying cell differentiation through the computational modeling of dynamical GRNs. Recent advances in cancer research have deviated their focus from the identification of cancer associated genetic mutations to the analysis of underlying complex GRNs to reach a mechanistic explanation for the emergence of cancer. In this chapter we review some advances in cancer modeling from the attractor landscape scheme, highlighting aspects of the disease that can be explained from this perspective. Our intention is to show the advantages of this systemic approach over a purely descriptive genetic approach, and its necessity to reach a mechanistic understanding of cancer.

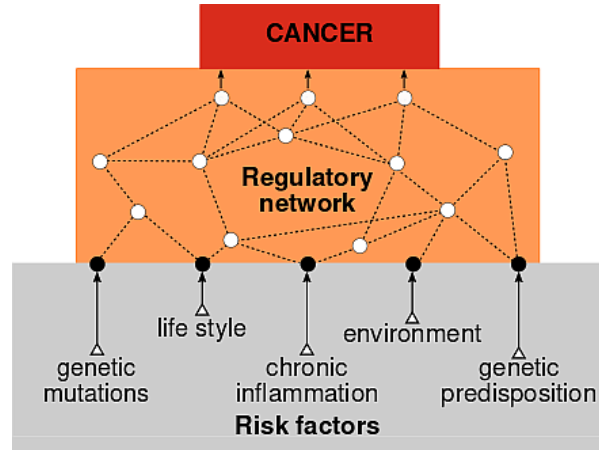
---

\*Correspondence: elenabuylla@protonmail.com

THE multiplicity of genetic, environmental and physiological factors involved in cancer appearance and progression makes its comprehension elusive through reductionist perspectives [1–3]. Nevertheless, mainstream cancer research still considers it a genetic-based disorder: a diverse group of diseases that result as a consequence of changes in the DNA [4]. This genocentric conception of cancer is reflected in reductionist approaches aiming at identifying genetic mutations as the causal factors for the disease [5, 6]. In spite of the genetic evidence associated with cancer, this approach has not been able to achieve a complete understanding of the disease. There is mountful evidence pointing at the necessity of a systemic perspective that departs from genetic mutations as the only explanation for the origin of cancer 1. For example: cells can become cancerous in the absence of mutations through trans- or dedifferentiation [7–9]; cancer cells manifest morphological and transcriptional convergence independently of the tissue of origin [10]; cancerous cells can be ‘normalized’ by several experimental non-genetic approaches [11–13]. These observations and the fact that carcinogenesis invariably recapitulates processes normally occurring during embryogenesis [14, 15], call for a developmental, rather than an entirely genetic, view of cancer.

A perspective on cancer coming from systems biology, seeks not to find the immediate molecular explanations for the appearance of a given kind of cancer, but to understand the generic mechanisms underlying cellular or tissue malignant transformation [2, 16]. Developmental transitions between cell types are a fundamental property of multicellular organisms, that can occur in the absence of genetic changes. A systemic approach implies that cancer is a developmental disease guided by the same mechanisms involved in cellular differentiation, that normally produce the diversity of cell types in multicellular organisms during development [2]. The epigenetic landscape proposed by Conrad H. Waddington is a scientific metaphor used to understand the regulatory constraints underlying development and cell differentiation [17]. From this perspective, the existence of multiple distinct phenotypic states (cell types) arising from clonal cell populations is explained by the dynamics of an underlying multistable gene regulatory network (GRN), as a complex dynamical system. This dynamical system is the mathematical formalization of the epigenetic landscape, and cancer is conceived as a special feature of it. This idea has already been proposed and developed by Stuart Kauffman in the 1970s, when he hypothesized that cancerous cells could be conceived as abnormal attractor states behaving like abnormal cell types [18]. This idea has been further expanded by Sui Huang, who defined cancer as a disease associated with the evolution of multicellularity, summarizing his idea with this phrase: “think of cancer as the price we pay for the capacity of evolving and developing a multicellular organism with one genome” [2].

The field of systems biology has developed a mechanistic methodology to study development and cell differentiation by building gene regulatory network (GRN) models from experimental evidence, and computationally simulating their



**Figure 1: Cancer as an emergent process.** A systemic approach to cancer must consider it as an emergent process from the interrelationship of genetic, environmental, and developmental processes.

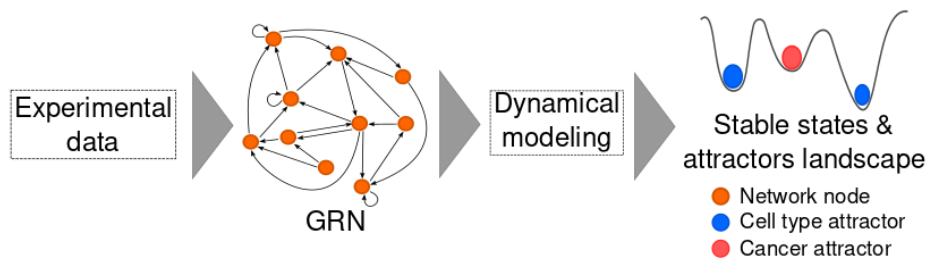
dynamics in order to reach the attractor states corresponding to observed cell types [19]. These network modeling approach has been already applied to the study of cancer, aiming to reveal the core regulatory mechanisms for its genesis and development, as well as to generate qualitatively different predictions from those coming from the somatic mutation theory [3]. This kind of research has already been undertaken by multiple research groups studying different aspects and kinds of cancers. In this chapter we pursue to make a review of this kind of models to have an overview of the current stage of knowledge in the GRN modeling for cancer.

#### GRN MODELING CAN EXPLAIN CELL TYPES AS ATTRACTOR STATES AND FORMALLY REPRESENT THE EPIGENETIC LANDSCAPE

Dynamical modeling of GRNs has become a well-established framework for the study of differentiation and cell type specification during development. In this framework, a GRN that represents mutual gene regulatory interactions is modeled as a multistable dynamical system. Given the nonlinear character of the GRNs, its dynamical behavior reaches different stable states, i.e. states where the regulatory constraints imposed by the network make the expression of each gene to stay unchanged [19]. Borrowing concepts from nonlinear dynamics, the stable stationary states are called attractors, and these states operationally correspond to configurations of gene expression or protein activation that underlie or correlate with cellular phenotypes. Dynamical modeling of GRNs can be done using either discrete algebra (e.g. Boolean or multi-valued logic) or a continuous approach using differential equations. Dynamic discrete models do not require kinetic pa-

rameters, which makes them more computationally feasible and allows them to be constructed using qualitative biological data. Regardless of the method used for their dynamical modeling, GRN models assume that the structure of the biologic networks they describe is more important than the kinetics of individual reactions and acquire their richness through the large number of interactions included in them.

GRN modeling has been extended not only to explain cell types as attractor states, but to formally represent the epigenetic landscape. The key for this formalization is to consider that, as well as generating the cellular phenotypic states (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 (basins of attraction), restricting the possible trajectories from one state to another one. In this context, the number, depth, width, and relative position of the basins of attraction would correspond to the hills and valleys of the metaphorical epigenetic landscape. For a more profound explanation of the methodology for GRN dynamical modeling and the inference of attractor epigenetic landscape refer to Davila-Velderrain *et al.* 2015 and references therein.



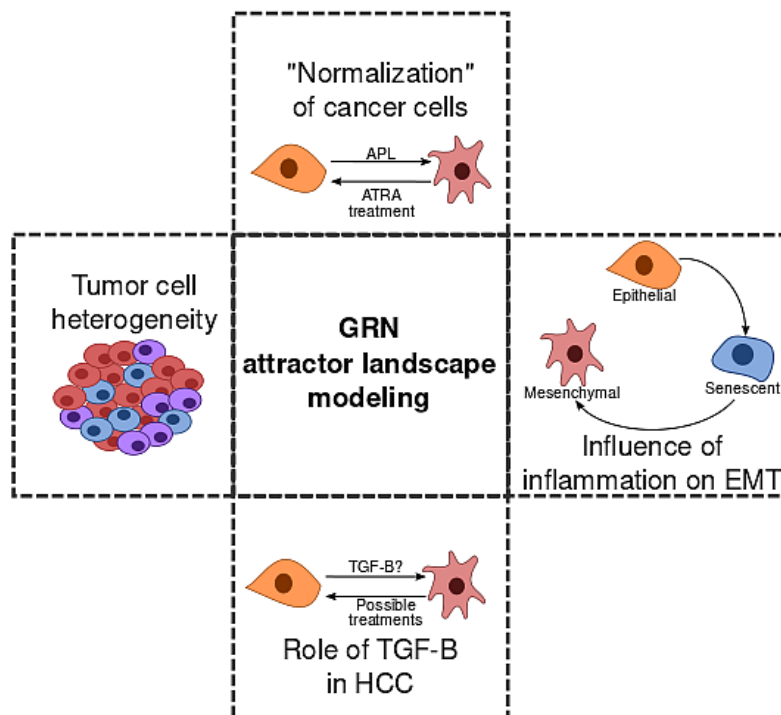
**Figure 2:** General framework for attractor landscape modelling of a cancerous process.

#### CANCER APPEARANCE THROUGH EPIGENETIC LANDSCAPE MODELLING

Here we will review some research papers tackling the problem of understanding cancer from the attractor landscape modeling methodology. In general, these methods propose a GRN based on experimental evidence, infer its associated epigenetic attractors landscape by exploring the network’s dynamical behavior, and from this model test some hypotheses related to cancer (Figure 2). We focus on different aspects of cancer that have controversial explanations from the mainstream genocentric approach and in which epigenetic landscape gives alternative explanations. Specifically, we focus on the cell heterogeneity in cancer tumors, an explanation for treatments that “normalize” cancer cells, and the spontaneous mutation free appearance of cancer and its association with chronic inflammation (Figure 3).

## Sources of cell heterogeneity in gastric cancer

Phenotypic and functional heterogeneity arising in tumoral cells is a shared feature of many types of cancer [20]. Mainstream cancer research offers two possible explanations for this phenomenon: clonal evolution [21] and cancer stem cell theory [22]. The clonal evolution hypothesis states that tumor heterogeneity is the result of heritable genetic and epigenetic variation, in other words heterogeneity comes from different mutations that appear in cancerous cells. The cancer stem cell hypothesis states that within a tumour there are cancer stem cells that give rise to various differentiated states.



**Figure 3:** Different phenomena with a complicated explanation from a genetic perspective can be explained by the analysis of GRNs underlying cancer.

Ao Ping's research group approached the question of cellular heterogeneity in cancer from the epigenetic landscape perspective. In particular, they addressed the issue of tumorous cell heterogeneity in gastric cancer by analyzing the attractor landscape of the associated GRN [23]. In their work, Li and collaborators follow the typical attractor landscape methodology. First they built a GRN including transcription factors, growth factors, cytokines, signal transduction pathways, and the interactions among them. They modeled the dynamics of such GRN using a

continuous approach, and found the stable states and possible transitions among them. The system's dynamic simulations reach 8 stable attractors and 13 saddle point states corresponding to transition states among the basins of attraction. They further validated the attractor states by modeling their system with Boolean discrete dynamics finding the same attractors, showing that the primary properties of the attractor landscape are defined by the network structure rather than by specific parameters. The attractors recovered represent 4 general cell states according to the activity of known molecular markers: cell cycle arrest (three attractors), proliferation (two attractors), cell death (one attractor), and stress response (two attractors).

Human gastric cancer cells have been classified into two phenotypes based on their gene expression: a gastric and an intestinal epithelial cell types [24]. Comparing their obtained attractor states with experimental expression data, they determined that cell cycle arrest states correspond to normal gastric epithelium and the proliferation states correspond to gastric cancer cells. The expression state of the two proliferating attractors corresponded to the gastric and intestinal types found in gastric cancer. These two attractors are maintained by two different feedback loops: the *Gastrin-Wnt/ $\beta$ -Catenin-Cdx2* loop and the *Sox2-SHH* loop, responsible for intestinal and gastric differentiation respectively. From this analysis, they showed the existence of two kinds of gastric cancerous cells. Furthermore, the multiple proliferative attractors recovered by the model can be explained by a regulatory mechanism intrinsic to the underlying GRN.

Expanding their network analysis, they looked for other possible sources of cell heterogeneity by looking for the possible paths a normal cell can take to arrive to the cancerous state. They explored this possibility by analyzing the transition routes from the normal gastric state to the two cancer attractors finding 16 different trajectories in the state space for transitions between attractors. A normal cell can pass through different attraction basins driven by non-genetic alterations, like fluctuations in gene expression or environmental noise. The existence of 16 different trajectories to cancer indicates that in the road to become cancerous, a gastric cell can pass through different transitory states and thus result in diverse cancerous cell states. As long as the phenotypic heterogeneity in the cancer cells does not affect the feedback loops that keep them in the cancer state, there can be heterogeneity among cancer cells.

In summary, Li and collaborators found two probable origins for gastric cancer heterogeneity, without the need to invoke de novo genetic mutations or cancer stem cells. It is important to highlight that this systemic explanation does not deny the appearance of new genetic mutations in cancer tumors. In fact, these alterations can be easily incorporated in the model but they are not necessary for the appearance of cancer nor its associated cellular heterogeneity.

### The effect of an efficient treatment for acute promyelocytic leukemia

Acute promyelocytic leukemia (APL) is a special type of leukemia because, unlike other types of leukemia, there is a therapy for treating it that “normalizes” leukemic blasts back to granulocytic differentiation. This therapy is based in treating leukemic patients with a combination of all-trans retinoic acid (ATRA) and arsenic trioxide (ATO) [25]. The genetic explanation for the origin of APL is the t(15;17) translocation, causing a *PML/RAR $\alpha$*  gene fusion [26]. Despite the existence of a genetic explanation for the disease and an efficient treatment for it, there are still open questions to have a complete understanding of what happens in APL (Yuan *et al.* 2016 and references there in).

With this concerns, Ao Ping’s group applied a dynamical network methodology to try to understand APL and how ATRA treatment causes its remission [27]. They built a regulatory network including molecules and molecular pathways critical for normal hematopoietic development and physiology. Their dynamical analyses found 18 attractors, which were classified in three groups according to their genetic expression patterns: proliferating-like attractors, differentiated-like attractors and attractors with apoptotic signatures. Among the proliferating-like attractors, they identified one attractor as a normal neutrophil progenitor and an APL-like attractor. They also identified in the differentiated-like attractors cell types of the hematopoietic hierarchy, according to their expression configuration [27].

After identifying APL and normal neutrophil attractors, they dynamically analyzed the network states around them to find possible trajectories in the attractor landscape to pass from a normal progenitor state to APL, and vice versa. In this way, the authors were able to identify the critical regulators mediating such attractor transitions. In order to pass from normal progenitor-like attractor to the APL-like one, it is necessary to induce a down-regulation in *BMP* signaling and upregulate *NRF2F2* and *SHH* signaling. On the contrary, passing from APL-like to normal progenitor-like is possible by down-regulating *SHH* and up-regulating *RUNX1* and *BMP* simultaneously; or by down-regulating *VEGF* and up-regulating *RARs*. These theoretical “normalizing” trajectories are consistent with the known effects of ATRA and ATO therapy, and constitute a mechanistic explanation of how this therapy works. In particular, ATO inhibits both *SHH* [28] and *VEGF* [29], while ATRA up-regulates *RARs* [30]; these effects are concordant with the activity changes found theoretically.

Their analysis also reveals that *SHH* and *NR2F2* are important molecular players in the maintenance of the APL cell state. Since *SHH* and *NR2F2* are important inducers of angiogenesis [31, 32], the authors propose that the APL-like attractor formation may be linked to angiogenesis. Angiogenesis is an important process during early embryonic development, so under this interpretation, APL might be considered as an erroneous reversion of an adult hematopoietic phenotype to an endothelial/mesenchymal one necessary during fetal development [27].

This work shows the existence of different network modules necessary for the attainment of an APL state, involving molecular pathways considered specific for embryonic organogenesis and mesenchymal development. In this way they ex-

pand the explanation for the origin of APL from a purely genetic basis caused by the t(15;17) translocation, to a systemic view in which APL is a regression to a fetal state necessary for angiogenesis. They are also able to show a probable mechanistic way of action of ATRA and ATO treatment to inhibit the normal-to-APL transition and enable an APL normalization, expanding the understanding of APL origin and treatment.

#### Spontaneous appearance of epithelial cancer

In a recent work from our research group, we aimed to find a core GRN underlying a conserved process observed in epithelial cell cultures *in vitro*, in which epithelial cells acquire first a senescent-like state that later evolves to a potential tumorigenic mesenchymal stem-like phenotype [33]. This process is characterized by a series of cell-state transitions, accompanied by the appearance of patterns of cellular promotion and progression, characteristics of epithelial carcinogenesis. We studied how spontaneous immortalization via EMT emerges from the regulatory interactions between molecular players with known contribution to the tumorigenic transformation of epithelial cells.

Our proposed network consist of a set of 41 molecular players (12 transcription factors and 29 signaling molecules) related to epithelial or mesenchymal cell differentiation, cellular inflammation, senescence, DNA damage, cell cycle, or epigenetic silencing; as well as 97 regulatory interactions between them. We analyzed the network dynamics with a Boolean approach and found that it converges to three stable attractors, corresponding to the epithelial, senescent and mesenchymal stem-like phenotypes according to their expression profiles. We tested 6 different mutant conditions (specifically, loss- and gain-of-function of *ESE-2*, *Snai2*, and *p16*) and show that our model is able to recover the experimentally grounded phenotypic consequences of these mutations.

After validating the model, we tested the effect of inflammation in EMT, as it has been recognized as one of the key drivers in carcinogenesis, partly due to its implication in EMT [34]. To do this, we simulated a forced activation of *NF- $\kappa$ B* node in the GRN and observed the changes in the attractors landscape. We found that cellular inflammation increased the size of the mesenchymal stem-like attractor basin from 56.25 to 75% while decreasing the region of convergence of the epithelial attractor (from 17.97 to 6.25%), and of the senescent one (from 25.78 to 18.75%). Thus, the model correctly recapitulates that cellular inflammation increases the probability of a cell to enter the mesenchymal stem-like attractor, and provides a mechanistic explanation for such increase.

Finally, we used our model to study the probable sequence of attractor attainment using the stochastic methodology proposed in by Alvarez-Buylla and collaborators [35]. This analysis indicates that, considering only the regulatory constraints of the GRN, the epigenetic attractors landscape is structured in such a way that the most probable flow for a population of cells starting in the epithelial phenotype is to transit to a senescent phenotype and then to a mesenchymal stem-like



phenotype, corresponding to the cell-state transitions observed *in vitro* [7–9].

Together with the analysis of the effects of inflammation in the epigenetic landscape, this model shows that although an epithelial cell can acquire a mesenchymal stem-like phenotype even under mutation-free, unperturbed physiological conditions, the likelihood of reaching this state is increased when pro-inflammatory conditions are present. Thus, providing a systems-level mechanistic explanation for the carcinogenic role of chronic inflammatory conditions [7, 36].

#### MODELING THE CANCEROUS EPIGENETIC LANDSCAPE CENTERED AROUND SPECIFIC GENES

GRN dynamical modeling also enables analyses to be focused on the effects that alterations on specific genes have on the epigenetic attractors landscape. In this section we will review two works that use network modeling and attractor landscape analyses to better understand the role of specific genes with important activities in cancer. We want to highlight these approaches because they show a way in which epigenetic attractor landscape modeling can incorporate and explain the role of mutations with known effects in cancer.

#### TGF- $\beta$ activity in hepatocellular carcinoma EMT

Epithelial-to-mesenchymal transition (EMT) is known to be a central process in cancer progression and metastasis [37]. In the previous section we revisited a study that found a mechanistic explanation for EMT in epithelial cells *in vitro*. Now we will review work from Reka Albert's research group in which they focus on the role of *transforming growth factor- $\beta$*  (TGF- $\beta$ ) in EMT, specifically in hepatocellular carcinoma (HCC) [38]. They center their work around this gene because TGF signaling is a conserved driver of EMT in epithelial cancer models [39].

To understand TGF- $\beta$ 's role in HCC, Steinway and collaborators built a network model incorporating growth factors, receptors, signal transductions proteins, and transcription factors involved in EMT, and used Boolean modeling to simulate its dynamics, focusing on the systems behavior upon TGF- $\beta$  activation [38]. An important detail of their model is that they take into account the different time scales of the interactions involved in their network, signal transduction events take seconds, while transcriptional events in minutes [40], by implementing a stochastic asynchronous updating scheme with a ranking system in their simulations [38]. Using this dynamical model, they find that their network reaches two attractor states corresponding to epithelial and mesenchymal phenotypes.

As mentioned above, their work is centered on the role of TGF- $\beta$  has in EMT. They validate this hypothesis by simulating its activation in the epithelial stable state causing the system to transit to the mesenchymal attractor. Since their dynamical model uses asynchronous stochastic update, after TGF- $\beta$  activation the system follows different transitory routes toward the mesenchymal state. Still, no matter what trajectory the system follows, it always reaches the mesenchymal

state, demonstrating that  $TGF-\beta$  activation is a sufficient condition for EMT. Their simulations also allow them to dissect the way  $TGF-\beta$  activation leads to a mesenchymal state, showing that it is driven by the joint activation of  $WNT$  and  $SHH$  signaling pathways. They validate their theoretical results testing the activation of the  $WNT$  and  $SHH$  pathways *in vitro*, measuring transcript levels of pathway markers after  $TGF-\beta$  induction in epithelial cell lines, confirming experimentally their computational results [38].

On a later work by the same group, they use their previously constructed HCC EMT network to identify molecular targets that could suppress  $TGF-\beta$  driven EMT [41]. Their network model allows them to test the systemic effect of thousands of individual and combinatorial node knockout perturbations, something that would hardly be possible experimentally, and measure their effects on the system behavior after  $TGF-\beta$  activation from the epithelial state. The knock down simulations are done by setting one or a combination of nodes permanently inactive and running the network dynamics. They test for all possible one, two, three and four-node combinations, giving them hundreds of thousands of possible combinations. Surprisingly, they only find 13 node combinations that inhibit EMT: seven single node and six combinations of two nodes knock downs. The seven single nodes correspond to the direct *E-cadherin* regulating transcription factors, an expected result given that loss of *E-cadherin* is widely considered a hallmark of EMT [42]. All six double-node combinations that inhibit EMT include the inhibition of *SMAD*, highlighting the importance of this protein in EMT but also the necessity of combinatorial interventions for its inhibition.

They also use their single knockout simulations to explore changes in the attractors landscape after node perturbations on the network that are not capable of inhibiting EMT. They show that knocking down single nodes that do not inhibit EMT in many cases causes the appearance of a new attractor intermediate between epithelial and mesenchymal phenotypes. The existence of these hybrid states with epithelial and mesenchymal features had already been reported in experimental models [43–45], but this network analysis gives an explanation for the appearance of these hybrid states in cancer. As with their previous work, they verify their results *in vitro* by testing their EMT inhibitory combinations with siRNAs in epithelial cell lines [41].

This systems biology approach was able to integrate available regulatory information to understand the mechanism through which  $TGF-\beta$  causes EMT in HCC and to identify ways to inhibit it. The large number of molecules and their possible combinations involved in EMT makes a thorough experimental screening for targets to inhibit EMT practically impossible. Alternatively, Albert's group demonstrate the utility of a computational dynamical systems approach to tackle this question and reach a testable set of candidate targets.

### Differences in *p53* network drive cellular fate choice before and after cancer

Another approach to the effects certain genes have on the epigenetic landscape is the one explored by Choi and collaborators as they studied the effects of DNA damage on the attractors landscape of a GRN centered around *p53* under normal or cancerous conditions. In their paper they are not trying to explore the origin of cancer, but instead the effects of a mutations associated to cancer in the underlying GRN dynamics. The authors assumed that *p53* has an important role in cancerous cell lines, and then analyzed the differences in the attractors landscape of a GRN with and without genetic alterations associated with breast cancer. In this sense, even though they are coming from a genocentric perspective, assuming genetic mutations are the cause of a cancerous state, they use the attractor landscape approach to understand why *p53* is an important player in cancer and its role in the network dynamics controlling cell fate [46].

Summarizing their results, they modeled the dynamics of a GRN module simulating *p53* activation by DNA damage. Afterwards, they modify the network incorporating alterations associated with breast cancer, modeled as up- and down-regulation of network nodes, and once again model the dynamics under DNA damage. The phenotypes they studied are the different behaviors a cell can undertake, which are: proliferation, cell cycle arrest, cell senescence, or cell death, and correspond to the different attractors of their epigenetic landscapes. They find that after DNA damage, *p53* activation makes normal cells enter a state of either cell death or cell cycle arrest, whereas cancer cells avoid entering cell death and instead stay in a senescent or cell cycle arrest state.

Through the state-space analysis, they not only elucidate the differential *p53* dynamics that modulate the cellular response to DNA damage, but also show that attractor landscape analysis can serve as a framework to identify the regulators that can be target of novel therapies. This is achieved by simulating alterations in the activity of different nodes of the network, and selecting those that make the system transit to the apoptotic attractor of the cancerous attractor landscape. It is important to mention that their analyses were coupled with experiments to empirically validate the predictions of the dynamical model [46].

## CONCLUSIONS

We examined different studies using an epigenetic attractors landscape modeling approach to cancer. This highlights the contribution of a systems biology approach to the understanding of cancer far from the genocentric view. Several evidences point to the necessity of leaving the mutational box to understand cancer and reach a wider and better understanding of the disease etiology and progression [2, 14, 34]. Following this idea, epigenetic attractor modeling of cancer is an opportunity to achieve a better insight on cancer and a way to understand it as a developmental disease, unchaining it from a solely genetic determinism.

As the examples presented above show, GRN modeling of cancerous processes

gives a formal interpretation of phenomena, like tumour heterogeneity and cancer normalization, that are difficult to explain from a gene-centric approach (Figure 3). Also, the GRN models can easily incorporate genetic mutations as one, but not the only or the most important, cause of cancer. Finally, they can be useful to propose therapeutic targets [41]. In this way, we underscore the importance of systems biology modeling approach to reach a better understanding of the disease, find ways to reduce its incidence in the population and find new treatments when cancer is already present.

## REFERENCES

1. Laubenbacher, R. *et al.* A Systems Biology View of Cancer. *Biochim Biophys Acta* **1796**, 129–139 (2010).
2. Huang, S. On the intrinsic inevitability of cancer: From foetal to fatal attraction. *Seminars in Cancer Biol.* **3**, 182–199 (2011).
3. Yuan, R., Zhu, X., Wang, G., Li, S. & Ao, P. Cancer as robust intrinsic state shaped by evolution: a key issues review. *Reports on Progress in Physics* **4** (2017).
4. Stratton, M., Campbell, P. & Futreal, P. The cancer genome. *Nature* **7239**, 719–724 (2009).
5. Network, T. C.G.A. R. *et al.* The cancer genome atlas pan-cancer analysis project. *Nat. Gen.* **10**, 1113–20 (2013).
6. Tomasetti, C., Li, L. & Vogelstein, B. Stem cell divisions, somatic mutations, cancer etiology, and cancer prevention. *Science*, 1330–1334 (2017).
7. Mani, S. *et al.* The epithelial-mesenchymal transition generates cells with properties of stem cells. *Cell* **4**, 704–15 (2008).
8. Xu, J., Lamouille, S. & Derynck, R. TGF- $\beta$ -induced epithelial to mesenchymal transition. *Cell* **2**, 156–172 (2009).
9. Li, C. *et al.* Epithelial-mesenchymal transition induced by TNF- $\alpha$  requires NF- $\kappa$ B-mediated transcriptional upregulation of Twist1. *Cancer Res.* **5**, 1290–300 (2012).
10. Ben-Porath, I. *et al.* An embryonic stem cell-like gene expression signature in poorly differentiated aggressive human tumors. *Nat. Gen.* **5**, 499–507 (2008).
11. Willhauck, M. *et al.* Reversion of tumor phenotype in surface transplants of skin SCC cells by scaffold-induced stroma modulation. *Carcinogenesis* **3**, 595–610 (2006).

12. Mahalingam D., K. C.L.J.T.L.Y.H.W. X. Reversal of aberrant cancer methylome and transcriptome upon direct reprogramming of lung cancer cells. *Sci. Rep.* **2** (2012).
13. Wang, Z. *et al.* Epigenetic reprogramming of human lung cancer cells with the extract of bovine parthenogenetic oocytes. *J. Cell and Mol. Med.* **9**, 1807–1815 (2014).
14. Micalizzi, D., Farabaugh, S. & Ford, H. Epithelial-mesenchymal transition in cancer: parallels between normal development and tumor progression. *J. of mammary gland biology and neoplasia* **2**, 117–134 (2010).
15. Kaufman, C. *et al.* A zebrafish melanoma model reveals emergence of neural crest identity during melanoma initiation. *Science* **6272** (2016).
16. Huang, S., Ernberg, I. & Kauffman, S. Cancer attractors: A systems view of tumors from a gene network dynamics and developmental perspective. *Seminars in Cell & Dev. Biol.* **7**, 869–876 (2009).
17. Waddington, C. *The Strategy of the Genes* (Allen and Unwin, 1957).
18. Kauffman, S. Differentiation of malignant to benign cells. *J.Theor Biol.* 429–451 (1971).
19. Davila-Velderrain, J., Martinez-Garcia, J. & Alvarez-Buylla, E. Modeling the epigenetic attractors landscape: toward a post-genomic mechanistic understanding of development. *Frontiers in genetics* **160** (2015).
20. Meacham, C. E. & Morrison, S. J. Tumor heterogeneity and cancer cell plasticity. *Nature* **7467**, 328–337 (2013).
21. Nowel, P. The clonal evolution of tumor cell populations. *Science*, 23–28 (1976).
22. Reya, T., Morrison, S., MF, C. & IL., W. Stem cells, cancer, and cancer stem cells. *Nature*, 105–111 (2001).
23. Li, S., Zhu, X., Liu, B., Wang, G. & Ao, P. Endogenous molecular network heterogeneity within gastric cancer. *Oncotarget* **15**, 13607–627 (2015).
24. Furihata, C. *et al.* Gastric- and intestinal-type properties of human gastric cancers transplanted into nude mice. *Cancer Res.* 727–733 (1984).
25. Wang, Z.-Y. & Chen, Z. Acute promyelocytic leukemia: from highly fatal to highly curable. *Blood*, 2505–2515 (2008).
26. Lo-Coco, F. & Hasan, S. K. Understanding the molecular pathogenesis of acute promyelocytic leukemia. *Best Pract. Res. Clin. Hematol.* 3–9 (2014).

27. Yuan, R., Zhu, X., Radich, J. P. & Ao, P. From molecular interaction to acute promyelocytic leukemia: Calculating leukemogenesis and remission from endogenous molecular-cellular network. *Sci. Rep.* **1** (2016).
28. Beauchamp, E. *et al.* Arsenic trioxide inhibits human cancer cell growth and tumor development in mice by blocking Hedgehog/ GLI pathway. *J. Clin. Invest.* 148–160 (2011).
29. Xiao Y.F. and Liu, S., Wu, D., Chen, X. & Ren, L. Inhibitory effect of arsenic trioxide on angiogenesis and expression of vascular endothelial growth factor in gastric cancer. *World J. Gastroenterol.* (2006).
30. Adamson, P. All-Trans-Retinoic Acid Pharmacology and Its Impact on the Treatment of Acute Promyelocytic Leukemia. *The Oncologist* **5**, 305–314 (1996).
31. Li, Y. *et al.* Sonic hedgehog (Shh) regulates the expression of angiogenic growth factors in oxygen-glucose-deprived astrocytes by mediating the nuclear receptor NR2F2. *Mol. Neurobiol.* 967–975 (2013).
32. Litchfield, L. M. & Klinge, C. M. Multiple roles of COUP-TFII in cancer initiation and progression. *J. Mol. Endocrinol.* 135–148 (2012).
33. Méndez-López, L. F. *et al.* Gene regulatory network underlying the immortalization of epithelial cells. *BMC Systems Biol.* **1**, 1–15 (2017).
34. Smith, M. & Peeper, D. Epithelial-mesenchymal transition and senescence: two cancer-related processes are crossing paths. *Aging* **10** (2010).
35. Alvarez-Buylla, E. *et al.* Floral morphogenesis: stochastic explorations of a gene network epigenetic landscape. *PloS One* **11**, e3626 (2008).
36. Ye, X. & Weinberg, R. Epithelial–mesenchymal plasticity: a central regulator of cancer progression. *Trends in Cell Biol.* **11**, 675–686 (2015).
37. Acloque, H., Adams, M., Fishwick, K., Bronner-Fraser, M. & Nieto, M. Epithelial–mesenchymal transitions: the importance of changing cell state in development and disease. *J Clin Invest.* 1438–49 (2009).
38. Steinway, S. N. *et al.* Network modeling of TGF signaling in hepatocellular carcinoma epithelial-to-mesenchymal transition reveals joint sonic hedgehog and Wnt pathway activation. *Cancer Res.* **21**, 5963–77 (2014).
39. Amin, R. & Mishra, L. Liver stem cells and TGF-Beta in hepatic carcinogenesis. *Gastrointest. Cancer Res.* 27–30 (2008).

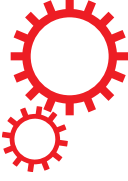
40. Papin, J., Hunter, T., Palsson, B. & Subramaniam, S. Reconstruction of cellular signalling networks and analysis of their properties. *Nat. Rev. Mol. Cell Biol.* 99–111 (2005).
41. Steinway, S. N. *et al.* Combinatorial interventions inhibit TGF $\beta$ -driven epithelial-to-mesenchymal transition and support hybrid cellular phenotypes. *NPJ Systems Biology and Applications* (2015).
42. Thiery, J. Epithelial–mesenchymal transitions in tumour progression. *Nat. Rev. Cancer*, 442–454 (2002).
43. Chao, Y., Wu, Q., Acquafondata, M., Dhir, R. & Wells, A. Partial mesenchymal to epithelial reverting transition in breast and prostate cancer metastases. *Cancer Microenviron.* 19–28 (2012).
44. Thomson, S. *et al.* A systems view of epithelial-mesenchymal transition signaling states. *Clin. Exp. Metastasis*, 137–155 (2010).
45. Jordan, N. V., Johnson, G. L. & Abell, A. N. Tracking the intermediate stages of epithelial-mesenchymal transition in epithelial stem cells and cancer. *Cell Cycle*, 2865–73 (2011).
46. Choi, M., Shi, J., Jung, S. H., Chen, X. & Cho, K.-H. Attractor Landscape Analysis Reveals Feedback Loops in the p53 Network That Control the Cellular Response to DNA Damage. *Science Signaling* **251** (2012).

# C Structural robustness of mammalian transcription factor networks reveals plasticity across development

Artículo publicado en 2018 en *Scientific Reports*.



# SCIENTIFIC REPORTS



OPEN

## Structural robustness of mammalian transcription factor networks reveals plasticity across development

J. L. Caldu-Primo<sup>1,2</sup>, E. R. Alvarez-Buylla<sup>1,2</sup> & J. Davila-Velderrain<sup>3,4</sup>

Network biology aims to understand cell behavior through the analysis of underlying complex biomolecular networks. Inference of condition-specific interaction networks from epigenomic data enables the characterization of the structural plasticity that regulatory networks can acquire in different tissues of the same organism. From this perspective, uncovering specific patterns of variation by comparing network structure among tissues could provide insights into systems-level mechanisms underlying cell behavior. Following this idea, here we propose an empirical framework to analyze mammalian tissue-specific networks, focusing on characterizing and contrasting their structure and behavior in response to perturbations. We structurally represent the state of the cell/tissue by condition specific transcription factor networks generated using DNase-seq chromatin accessibility data, and we profile their systems behavior in terms of the structural robustness against random and directed perturbations. Using this framework, we unveil the structural heterogeneity existing among tissues at different levels of differentiation. We uncover a novel and conserved systems property of regulatory networks underlying embryonic stem cells (ESCs): in contrast to terminally differentiated tissues, the promiscuous regulatory connectivity of ESCs produces a globally homogeneous network resulting in increased structural robustness. We show that this property is associated with a more permissive, less restrictive chromatin accessibility state in ESCs. Possible biological consequences of this property are discussed.

A central tenet of systems biology is that cell behavior can be understood in terms of the structure and dynamics of underlying complex molecular networks<sup>1,2</sup>. Under such paradigm, major efforts have been made to systematically map and characterize the properties of molecular networks at different levels of organization. Reference protein-protein interaction, metabolic, and transcriptional regulatory networks have been constructed and are being frequently updated in several model organisms<sup>3-5</sup>. Initial efforts have largely focused on providing an organismal reference for the global network structure.

Network theory provides methods for the systemic description of a network's structure and its dynamics<sup>6-8</sup>. One of the major results of network biology is the discovery within the reference networks of apparently universal organizational properties across the different types of complex biological networks<sup>2</sup>. While the characterization of reference real-world complex networks has uncovered structural similarities among complex networks that are believed to underly their systemic properties<sup>2,6</sup>, much less is known about the degree of structural heterogeneity of condition-specific biomolecular networks, and how patterns of variation promote or constrain systems-level behaviors.

In cell biology, one intriguing hypothesis is that network heterogeneity emanating from the normal process of development might result in differential behaviors underlying the contrasting cellular phenotypes. In line with this idea, the field of network biology has recently started shifting towards the characterization of condition-specific networks and analysis of circuitry dynamics<sup>9,10</sup>, presumably due to the increasing availability

<sup>1</sup>Centro de Ciencias de la Complejidad (C3), Universidad Nacional Autónoma de México, Cd. Universitaria, México, D.F., 04510, Mexico. <sup>2</sup>Instituto de Ecología, Universidad Nacional Autónoma de México, Cd. Universitaria, México, D.F., 04510, Mexico. <sup>3</sup>MIT Computer Science and Artificial Intelligence Laboratory, Cambridge, Massachusetts, USA. <sup>4</sup>Present address: Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA. Correspondence and requests for materials should be addressed to J.D.-V. (email: [jdavilav@mit.edu](mailto:jdavilav@mit.edu))

of functional genomics and epigenomics assays. For example, Neph and collaborators put forward a methodology to assemble tissue-specific transcription factor networks with the aid of available chromatin accessibility profiles from multicellular genomes<sup>9,11–13</sup>. The proposed networks connect each transcription factor (TF) to its incoming TF regulators, thus representing the regulatory structure of the cell in terms of the main regulators (e.g. TFs) and the mutual regulatory interactions among them. More specifically, using digital genomic footprinting (DGF) analysis, TF-TF interactions are established by integrating TF motif matching with DNase I hypersensitive sites (DHS) and high-resolution genomic footprints. Tissue-specificity comes from the condition-specific accessibility of cis-regulatory regions upstream a TF. Using this approach, tissue-specific TF networks have been constructed for model organisms and for human<sup>9,14</sup>. Given that the observed TF interactions reflect tissue-specific activity states, we reasoned that the structure and relative systems-level behavior displayed by these networks could provide insights into the biology and differentiation potential of the corresponding tissues.

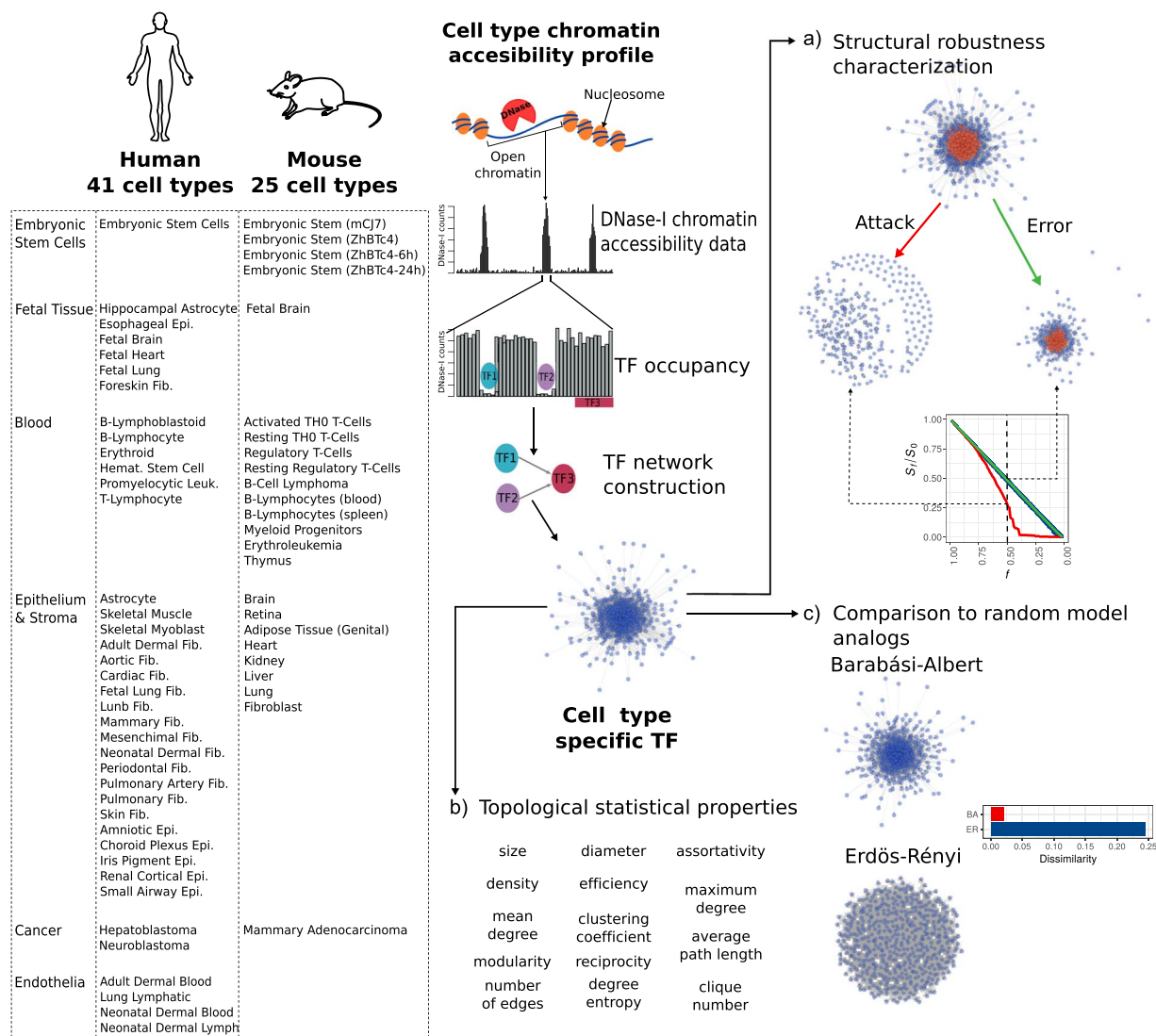
In order to begin understanding the link between network structure heterogeneity, behavior, and biological phenotypes, here we put forward a computational framework to characterize the structural properties of mammalian tissue-specific TF networks and their behavior, emphasizing the degree of deviation from theoretical expectations. We focus on one systems-level behavior which is informative of the latter: the robustness of the networks to structural perturbations. We profiled the structural properties of a broad set of TF networks in mouse and human, and we compared the observed behavior across tissues and against expectations of theoretical models. Interestingly, we discovered that embryonic stem cells (ESCs) possess a distinctive regulatory structure: its higher structural similarity to the topological properties expected from a homogeneous network theoretical model endows them with a remarkable resilient behavior. We show by analysing chromatin accessibility profiles, that the tissue-specific TF network captures at a systems level, the more permissive and less restrictive property of the ESC epigenome relative to adult, differentiated tissues. However, unlike previous studies quantifying developmental potential with a gene expression-based network entropy framework<sup>15,16</sup>, we did not find a robust distinction between adult stem and differentiated cell populations; which might indicate a limitation of the degree of resolution captured by TF networks and, consequently, of the structural robustness measure proposed here. We discuss potential biological implications, and future extensions.

## Results

**Analysis framework.** Networks provide a theoretical framework that allows a convenient conceptual representation of interrelations among a large number of elements<sup>6</sup>. Furthermore, it is usually possible to frame questions about the behavior of the underlying real system by applying well-established analyses on the network representing empirical data<sup>17</sup>. Here we focus on tissue-specific networks where nodes represent TFs and links inter-regulatory interactions, and propose an analysis framework with the goal of characterizing the commonalities and differences in behavior against structural perturbations across tissues. We ask whether some tissues display extreme behaviors, and whether or not such deviations and extreme behaviors highlight aspects of the underlying biology. We hypothesize that the differences to be discovered underlie aspects of the observed biological functionality and of the broad degree of differentiation of the tissues. The proposed framework includes the following steps (see Fig. 1). (1) The state of the cell is structurally represented by tissue-specific networks of regulatory interactions among transcription factors as proposed in<sup>9,14</sup>. Briefly, a TF is considered regulator of another TF when a motif instance of the former TF occurs within a DNase I footprint contained in the proximal regulatory region of the latter TF (10 kb interval centered on the transcription start site [TSS]). (2) The system's behavior of a network is defined as the response of the network against increasing structural perturbations<sup>2,18</sup>, and the response is measured by two metrics: the change in giant component size, and the change in efficiency, both relative to the original, unperturbed network (see Methods). The complete behavior is captured by the qualitative properties of the change from start until complete disruption; we introduce a simple metric to quantify it (Fig. 1a). (3) The structure of each network is numerically characterized by 14 topological measures (Fig. 1b). (4) The degree of deviation of each network relative to expectations from homogeneous (Erdős-Rényi) and heterogeneous (Barabási-Albert) random graph models is quantified (Fig. 1c).

After applying these steps to each tissue-specific network, we rank the networks based on the robustness of their behavior, we identify those displaying the most extreme response, and we statistically explain the behavior in terms of predictive topological features and relative deviation from analogous homogeneous and heterogeneous random models. Thus, starting from an input set of tissue-specific networks, our framework produces a structural robustness ranking, a set of structural features underlying the behavior, and a mapping of the networks into the homogeneous-heterogeneous network space.

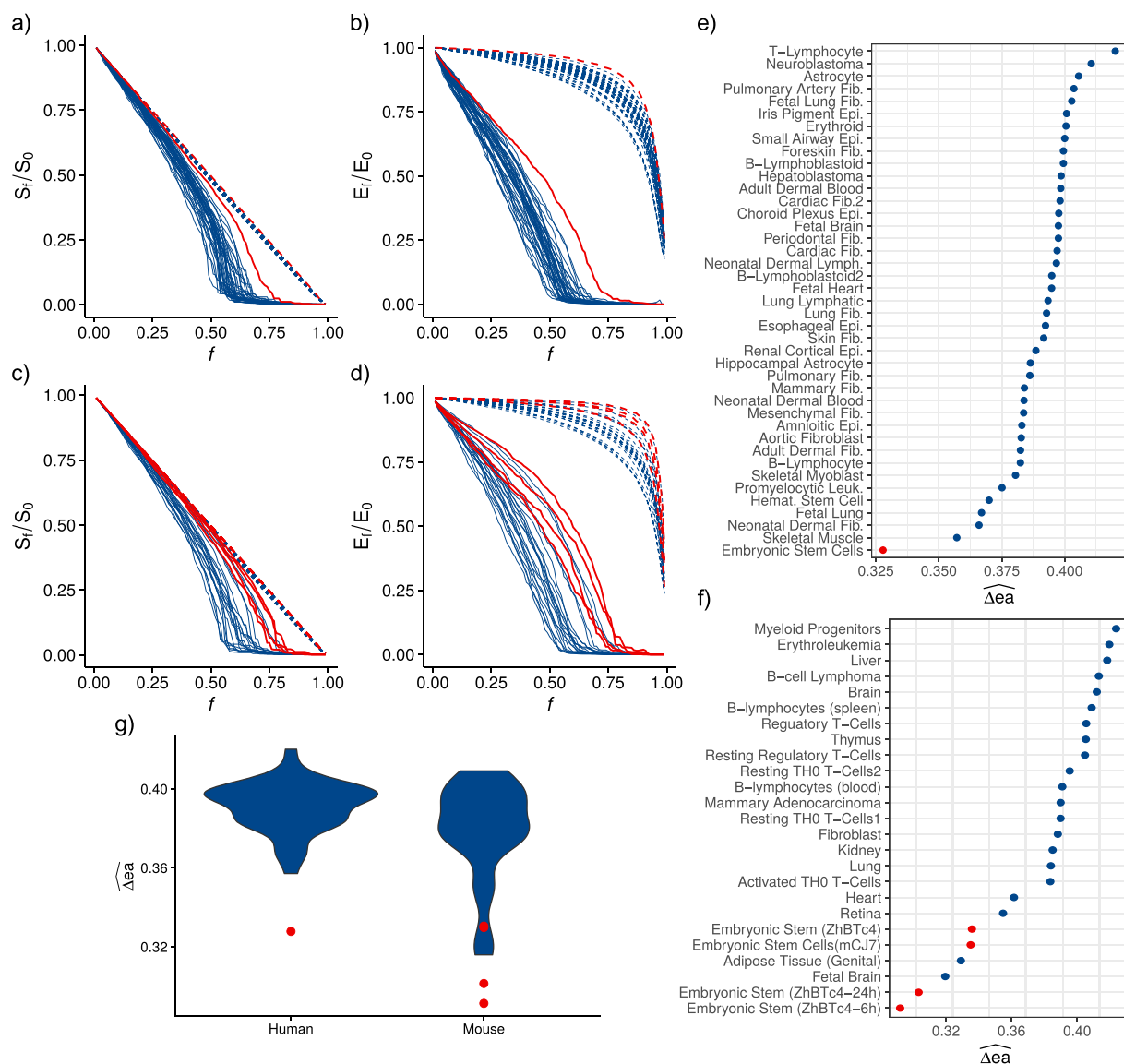
**Network structural differences reveal plasticity of systems behavior upon perturbation.** It has been shown that a differential response to random structural perturbations (errors) and directed alterations (attacks) enables a concrete distinction between homogeneous and heterogeneous networks in terms of systems' behavior<sup>18</sup>. A network representing a real complex system is expected to tolerate random failures, but to be more vulnerable against directed attacks targeting key, connected components. Taking this well-established framework, we evaluated the robustness behavior of TF networks across tissues. The operational definition of structural robustness applied here is based on an intuitive idea: disabling a substantial number of nodes will result in an inevitable functional disintegration of a network<sup>2</sup>, but the degree of tolerance will vary across tissues. We measured tolerance to random perturbations by randomly removing nodes from the networks and quantifying the change in the size of the largest connected component (giant component), and the change in network efficiency – an approximation to loss or gain of network connectivity (see Methods). For directed attacks, we repeated the experiments but sequentially removing nodes in decreasing order of centrality (degree) (Fig. 1a). We profiled the response to perturbations in 41 human and 25 mouse tissue networks.



**Figure 1.** Structural profiling of cell type specific TF networks. **(a)** Structural robustness was measured simulating attacks (removing high degree nodes (red)) and errors (removing randomly selected nodes). **(b)** Networks characterization was done measuring topological features of every network. **(c)** Each network was compared to random model networks by measuring its dissimilarity to an analogous ensemble of homogeneous and scale-free networks.

Overall, all networks were found to be highly tolerant to random errors. In both mouse and human tissues, the size of the giant component ( $S_f/S_0$ ) decreases linearly with  $f$  without abrupt transitions (Fig. 2a,c, dashed lines). The efficiency of the networks ( $E_f/E_0$ ) also shows consistent behavior across all human and mouse tissues: it shows minimal decrease for a large proportion of  $f$  until it falls abruptly around  $f=0.8$  (Fig. 2b,d, dashed lines). The observed robustness to random failures is consistent with predictions from percolation theory in complex random networks, as it is less likely to perturb key, highly connected components in networks with long-tail degree distribution<sup>6,18</sup>. Also consistent with theory, TF networks were found to be much more vulnerable to directed attacks. Interestingly, however, we observed a high degree of variability in the behavior upon attacks across networks. Both measures (giant component size and efficiency) revealed transitions at different fractions  $f$  of attacked nodes (see Fig. 2a–d, solid lines). Interestingly, we found that in both human and mouse the TF networks of embryonic stem cells (ESCs) display, relative to differentiated tissues, an extremely robust behavior against both failure and attack perturbations, the latter being much more pronounced (see Fig. 2a–d, red lines).

With the goal of quantitatively describing and to analyze the discovered patterns of heterogeneity among tissues, we define the metric error-attack deviation ( $\Delta ea$ ), which simply quantifies the degree of deviation of a given network's behavior upon directed attack perturbations from that stemming from random errors. We use this metric here as a measure of the structural robustness of complex networks to perturbations, as it reflects the degree to which attacks and errors are tolerated (see Methods). Intuitively, the smaller the value of  $\Delta ea$  the closer the global response of the network against attacks relative to that against error, indicating a higher degree of robustness. We performed the calculation individually for the two damage measures used in this study:  $S_f/S_0$  and



**Figure 2.** TF networks structural robustness. The behavior against errors (dashed) and attacks (solid) of every cell type is shown, red lines correspond to the ESCs behavior and blue lines to other cell types. **(a)** Human giant component size decrease. **(b)** Human efficiency decrease. **(c)** Mouse giant component size decrease. **(d)** Mouse efficiency decrease. **(e)** Human and **(f)** mouse vulnerability measure for each TF network. **(g)** Human and mouse vulnerability measures distribution, red dots correspond to ESC measurements.

$E_i/E_0$  (Supplementary Fig. 1). From these error-attack deviation measures, we defined network structural vulnerability ( $\Delta ea$ ) as the mean  $\Delta ea$  for giant component size and efficiency (see Methods). This measure enables the quantification of differential structural robustness to attacks displayed by the networks (cell types). The vulnerability measure of human and mouse cell types corroborates the heterogeneity of structural robustness among cell types, and the extremely deviating behavior of ESCs (Fig. 2). ESCs have an error-attack deviation significantly lower than other cell types, highlighting their significantly higher robustness against attacks relative to more differentiated tissues.

**Network structural rearrangement during differentiation.** The observed differences in structural robustness among tissues point to the existence of patterns of variation in global network structure. In order to characterize the structural heterogeneity of TF networks, we analyzed their topology and asked whether specific topological features more predominantly explain the observed robustness patterns. In particular, what structural features underlie the extreme robust behavior of ESCs? As a first approximation we simply asked how similar are networks among each other? We computed pair-wise dissimilarity scores for every pair of TF networks in mouse and human, using a structural dissimilarity ( $D$ ) approach (see Methods). Network dissimilarity is a useful method for network comparison as it quantifies structural topological differences based on node distance probability distributions, capturing nontrivial structural differences<sup>19</sup> – as opposed to the intuitive counting of presence or absence of common links.

Despite the fact that all TF networks are relatively similar – having average  $D$  values of 0.040 and 0.064 in human and mouse, respectively – there is variation in the structural similarity among them.  $D$  ranges from 0.003 to 0.160 in human, and from 0.003 to 0.184 in mouse. Considering pair-wise comparisons in human networks, ESC is the most dissimilar network for 24 (58.5%) of the tissues. For the remaining 17 tissues, the most dissimilar network corresponds to Astrocyte. These two tissues also have the highest  $D$  median scores: ESC (0.090) and Astrocyte (0.077). Interestingly, these two networks are also the most dissimilar between one another. Thus, the undifferentiated ESC localizes at one extreme of the topological space while the highly differentiated Astrocyte localizes at the other. We built a dendrogram using network dissimilarity as distance measure among human networks. ESC is clearly different from the other tissues as it is placed in a single branch at the bottom of the distance dendrogram, separated from all the other cell types (Fig. 3a).

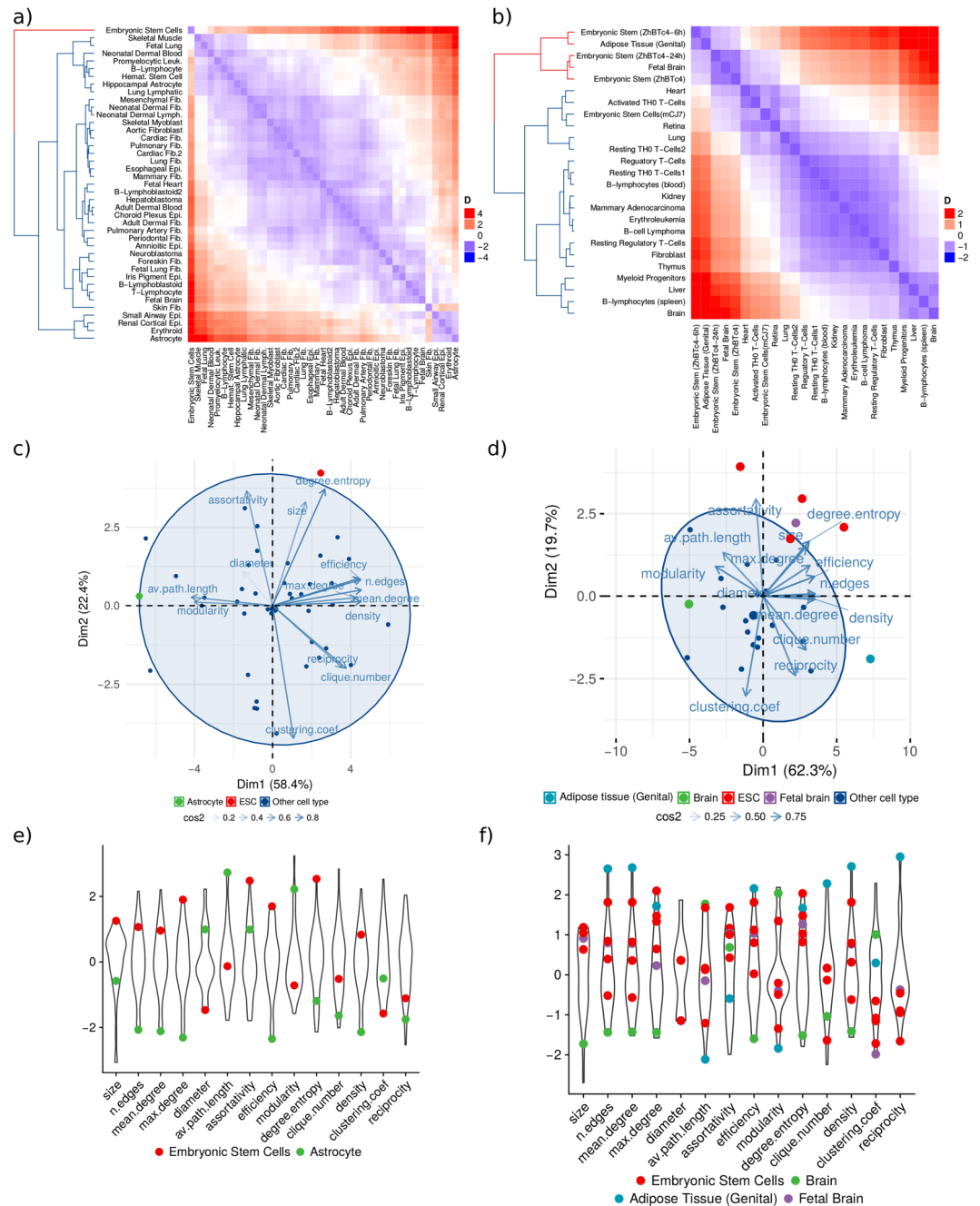
Mouse networks show a similar pattern to that found in human cell types. Pair-wise comparisons show that the most dissimilar networks are ESCs and the highly differentiated Brain, with these two tissues occupying the extremes in the dissimilarity distribution (Fig. 3b). ESC ZhBTc-6h has the highest  $D$  value for 16 of the 25 cell types (64%), while the other two ZhBTc ESCs also rank among the most different networks, and in the remaining 9 cell types the highest  $D$  value corresponds to Brain. Unbiased hierarchical clustering aggregates three ESC lines (ZhBTc, ZhBTc-6h, and ZhBTc-24h) in a separate basal branch, together with Genital Adipose Tissue and Fetal Brain. Fetal tissues are expected to display some degree of similarity with ESCs, due to overlap of developmental processes during fetal development. Adipose is an heterogeneous tissue, possibly including undifferentiated adipose stem cells. Overall, the topology of ESCs networks in mouse and human is clearly distinct from adult differentiated tissues such as brain and liver. In both dendrograms, differentiated tissues do not seem to be structured according to their lineage. This reflects that from a structural point of view, developmental lineages networks are not clearly distinguished and only a significant difference between ESCs and adult differentiated tissues is observed (Fig. 3). We reasoned that this observation might stem from a distinctive chromatin accessibility state characterizing ESCs, which we explore below. Overall, there is a significantly higher dissimilarity between ESCs and adult cell types than among those differentiated tissues (Supplementary Fig. 2).

To further explore the topological differences among tissues, we characterized the structure of every network using 14 standard measures for network topology description (Table 1, see Methods)<sup>6,7</sup>. These measures capture important characteristics of a network's global structure, which in part determines its functionality. In particular, we seek to dissect the structural heterogeneity among tissues, identify features associated with the observed robustness, and finally map those structural features that discriminate ESCs' networks from those of differentiated tissues.

We performed principal components analysis (PCA) using the measured topological features, in order to explore network aggregation behavior in the feature space, while at the same time avoiding collinearity. For both human and mouse data, the features with highest contribution for the first principal component (PC) are mean degree, number of edges, density, efficiency, and modularity. The former four features are highly correlated, all of them measuring network degree of connectivity. In spite of mean degree's high contribution to the first PC, this property does not explain the structural difference observed in ESCs: mean degree of ESCs does not deviate from the empirical distribution among other tissues (Supplementary Fig. 3, and Fig. 3e,f). The features contributing to the second PC are clustering coefficient, assortativity, and degree entropy. Projecting the networks to a 2D space based on PCs, we found no apparent clustering (Fig. 3c,d). However, a closer examination shows that, as expected, ESCs are separated from the other tissues, having higher values for the second PC. The highly specialized networks of Astrocyte and Brain tissue localize at the opposite extreme, evidencing the extreme structural differences relative to ESCs. These differentiated networks are characterized for having extremely low values for the first PC. Considering these patterns, ESCs are characterized for having high values of degree entropy and assortativity, but low clustering coefficient. On the other hand, Brain and Astrocyte networks have high modularity and average path length, but small density, efficiency, and mean degree. This pattern is confirmed by the features distribution (Fig. 3e,f).

The topological characterization corroborates an extreme difference in network topology between undifferentiated ESCs and differentiated tissues. Analysis of features distribution shows that tissues spread through a feature space following two main axes, one going from highly modular to highly efficient networks, and another separating highly degree entropic and degree assortative structures from those with high global clustering. ESCs are distinguished from differentiated tissues for having more interacting TFs, and these are globally connected in a more promiscuous way, as evidenced by higher levels of entropy in the degree distribution. In contrast, differentiated networks of Brain and Astrocyte are more structured, as evidenced by high levels of modularity, yet low levels of efficiency and density. Taking into account the existence of a trade-off between network efficiency and modularity<sup>20</sup>, this observation hints to a possible path of developmental dynamics of TF network structure in which the system transits from a configuration promoting efficiency in information flow and robustness, into a highly modular topology suggestive of functional specialization.

**Interpretation in terms of theoretical network models.** As mentioned above, robustness to directed attacks has been linked to homogeneous network topologies, in contrast to the “robust yet fragile” behavior characteristic of heterogeneous (scale-free) networks<sup>18</sup>. Considering this result, we compared each TF network to analogous ensembles of random homogeneous and scale-free networks generated using the Erdős-Rényi (ER) and the Barabási-Albert (BA) models, respectively (see Methods). ER networks with high number of nodes approach a Poisson degree distribution, symmetric for relatively high average degrees. On the contrary, BA networks have a characteristic right skewed power-law degree distribution. We compared the real world networks with the theoretical models, with the goal of placing them within a heterogeneity axis by quantifying deviations. Given the discovered high robustness to directed attacks and high degree entropy of ESCs, we reasoned that such a contrast will help clarify the global structural features underlying such behavior.



**Figure 3.** Networks structural profiling. **(a,b)** Dissimilarity among cell types, heatmaps of scaled  $D$  values among **(a)** human and **(b)** mouse cell types. **(c)** Human and **(d)** mouse networks topological features PCA. **(e)** Human and **(f)** mouse topological features distribution, colored dots show the value for each feature of the indicated cell type.

We measured network structural dissimilarity between each network and its ER ( $D_{ER}$ ) and BA equivalents ( $D_{BA}$ ). As expected, all networks are significantly more similar to BA than to ER networks.  $D_{ER}$  ranges from 0.191 to 0.285 and from 0.139 to 0.287; whereas  $D_{BA}$  ranges from 0.019 to 0.047 and from 0.013 to 0.055, in human and mouse respectively (Fig. 4c). The fact that BA networks are more similar to the TF networks is consistent with discoveries of other real world complex networks having scale-free topologies<sup>21</sup>. Interestingly, however, we found clear differences among the networks regarding their relative similarity to each theoretical model. For instance, ESCs have the lowest  $D_{ER}$  in both human and mouse (Fig. 4). In the case of  $D_{BA}$ , a contrasting pattern emerges: ESCs are among the tissues with higher values. Nevertheless, ESC  $D_{BA}$  values are not significantly different from those of other tissues, falling within the observed distribution of  $D_{BA}$  (Fig. 4c). Considering  $D_{ER}$  and  $D_{BA}$  together and taking both human and mouse networks, ESCs are separated from the other cell types, as shown in Fig. 4. A conserved pattern in both human and mouse emerges in which ESCs have a relatively lower dissimilarity to ER

Feature	Human			Mouse		
	Range	Average	Standard Deviation	Range	Average	Standard Deviation
Size	[493, 533]	521	9.27	[555, 583]	574.4	7.21
No. of edges	[9099, 18906]	14002.97	2272.36	[15392, 36448]	22970.4	5084.83
Mean degree	[36.03, 72.02]	53.65	8.33	[54.10, 125.036]	79.816	16.881
Diameter	[5, 8]	6.19	0.81	[5, 7]	5.76	0.66
Density	[0.034, 0.069]	0.052	0.008	[0.48, 0.10]	0.07	0.014
Average path length	[2.30, 2.69]	2.45	0.086	[2.08, 2.41]	2.26	0.086
Clique number	[14, 27]	19.39	2.68	[19, 34]	26.44	3.31
Clustering Coefficient	[0.24, 0.39]	0.29	0.036	[0.25, 0.37]	0.30	0.027
Assortativity	[-0.21, -0.12]	-0.17	0.022	[-0.18, -0.13]	-0.15	0.015
Efficiency	[0.47, 0.54]	0.51	0.017	[0.50, 0.59]	0.54	0.024
Modularity	[0.10, 0.17]	0.12	0.014	[0.06, 0.11]	0.09	0.013
Degree Entropy	[5.69, 5.94]	5.80	0.052	[5.79, 6.09]	5.93	0.078
Reciprocity	[0.03, 0.06]	0.05	0.006	[0.04, 0.09]	0.06	0.01
Maximum degree	[266, 416]	348.2	35.6	[334, 574]	436.9	65.2

**Table 1.** Topological features measured for every human and mouse network.

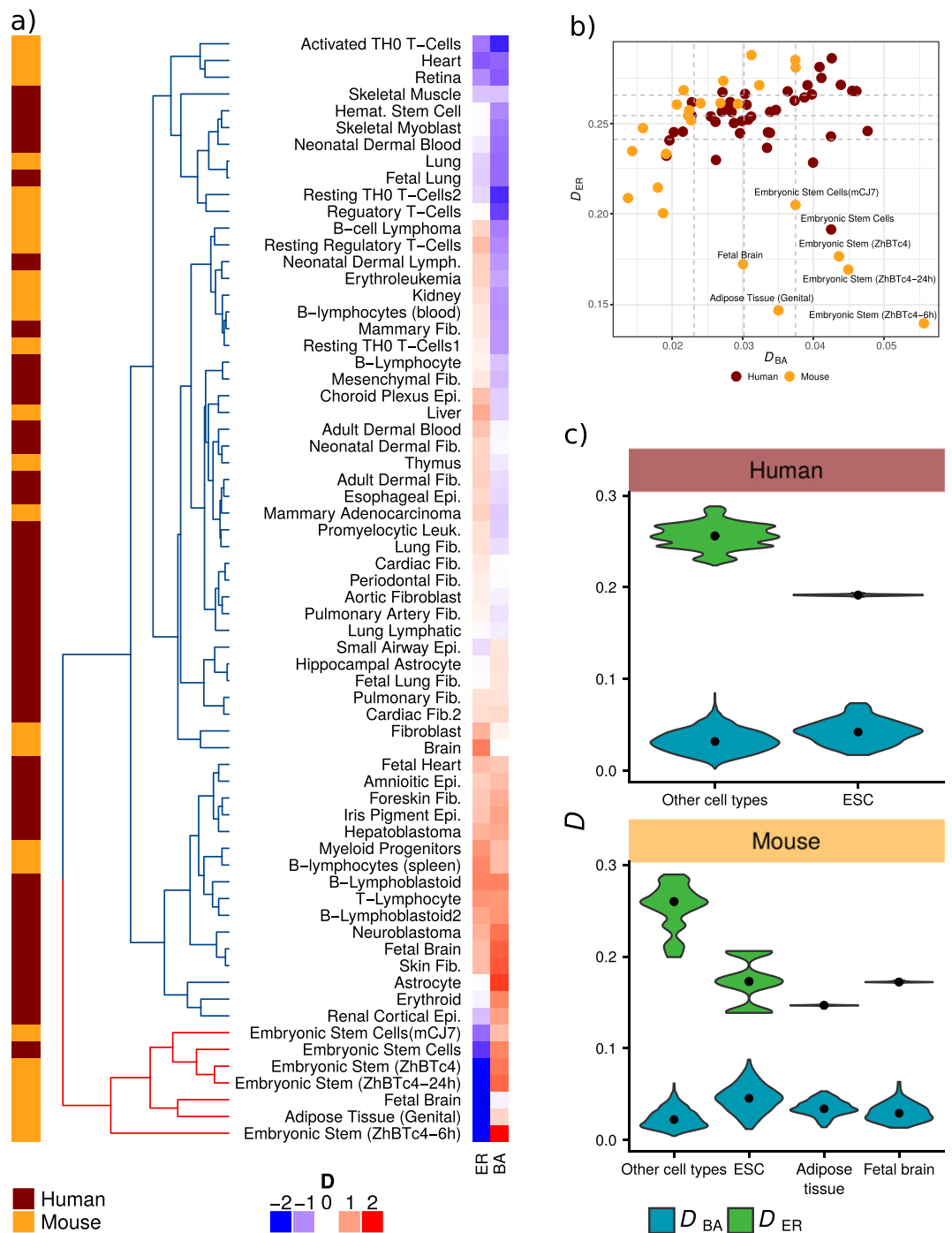
networks and a relatively higher dissimilarity to BA than the other tissues. As with the dendrogram created from  $D$  among networks (Fig. 3b), dissimilarity to model networks does not recover lineage hierarchies of differentiated tissues, yet it underscores a broad difference in global structure between ESCs and differentiated tissues.

For every model network we measured the same 14 topological properties we used to characterize cell type networks, and performed a PCA of their features including the real and model networks. In both human and mouse networks, the PCA graph shows a common pattern. The first component separates three clusters corresponding to each model network and the real networks, situating BA and ER networks in the extremes and the real networks between them, closer to the BA cluster (Fig. 5a,b). As shown in the structural dissimilarity analysis, this pattern confirms that real networks are more similar to scale-free networks than homogeneous networks. Real TF networks are situated in between BA and ER clusters, thus creating a feature space between the two model networks in which real networks can be situated. The pattern shows that ER networks tend to have higher degree entropy and assortativity, while BA networks tend to have higher diameter and clustering coefficient (Fig. 5a,b).

We show that ESC networks have a distinctly higher robustness against directed attacks relative to differentiated tissues. Since scale-free topology explains the fragility against directed attacks in complex networks, we analyzed the topological features of ESCs networks that deviate from BA expectations. We calculated the deviation of every real network feature compared to its distribution in the corresponding BA model (see Methods). From this analysis we selected features in which the real networks differ significantly from BA models, these are: average path length, assortativity, degree entropy, maximum degree, modularity, clique number, and clustering coefficient (Fig. 5c). We found extremely high deviation (z-score) on degree entropy, assortativity, and average path length in ESCs (Fig. 5d). From our PCA analysis, we know that these features have a high contribution to the first PC; in particular, ER networks tend to have higher degree entropy and assortativity. This indicates that, even though ESC networks are closer to a BA topology, the features for which they are different from a BA model are characteristic of ER networks. This is illustrated by visual contrast of ER expected and empirically observed values of the deviating features among cell types (Fig. 5f,g). Thus, we conclude that ESCs have extreme values in features characteristic of ER networks.

**Network homogeneity predicts structural robustness.** We show that the topological plasticity of tissue-specific TF networks can be characterized by comparing them to model networks. As mentioned before, this structural differences are associated with the networks' response to random and directed perturbations. To further understand the structural features underlying the observed structural robustness pattern, we fitted statistical models in an attempt to further uncover explanatory topological features. Using the previously defined network vulnerability ( $\widehat{\Delta ea}$ ) as the response variable, we fitted two statistical models: linear regressions using the 14 network features as well as  $D_{ER}$  and  $D_{BA}$  as predictors, and a random forest regression using the 14 topological features as predictors. For each model we measured its mean square error, and validated its accuracy through five-fold cross validation. The best predictor of network vulnerability is network's  $D_{ER}$  with a cross validation mean square error of 0.00022. There is a positive relationship between  $D_{ER}$  and  $\widehat{\Delta ea}$  (Fig. 6b), indicating that the more a network resembles a homogeneous network, the higher its structural robustness. The topological feature with the best predictive performance is degree entropy, a feature correlated with a network similarity to a homogeneous network. Thus, the deviation from ER model expectation  $D_{ER}$ , a measure quantifying the degree of homogeneity of a real-world network, and which is distinctively high in ESCs; is predictive of structural robustness.

**ESCs TF network structure captures a more accessible and permissive chromatin state.** Network structural analyses show that ESCs have a distinct network topology, mainly characterized by a higher uniformity in the number of interacting partners (degree entropy). Since the networks we analyzed reflect both the presence of a TF motif and DNase-seq chromatin accessibility signal<sup>9,14</sup>, we reasoned that at the global

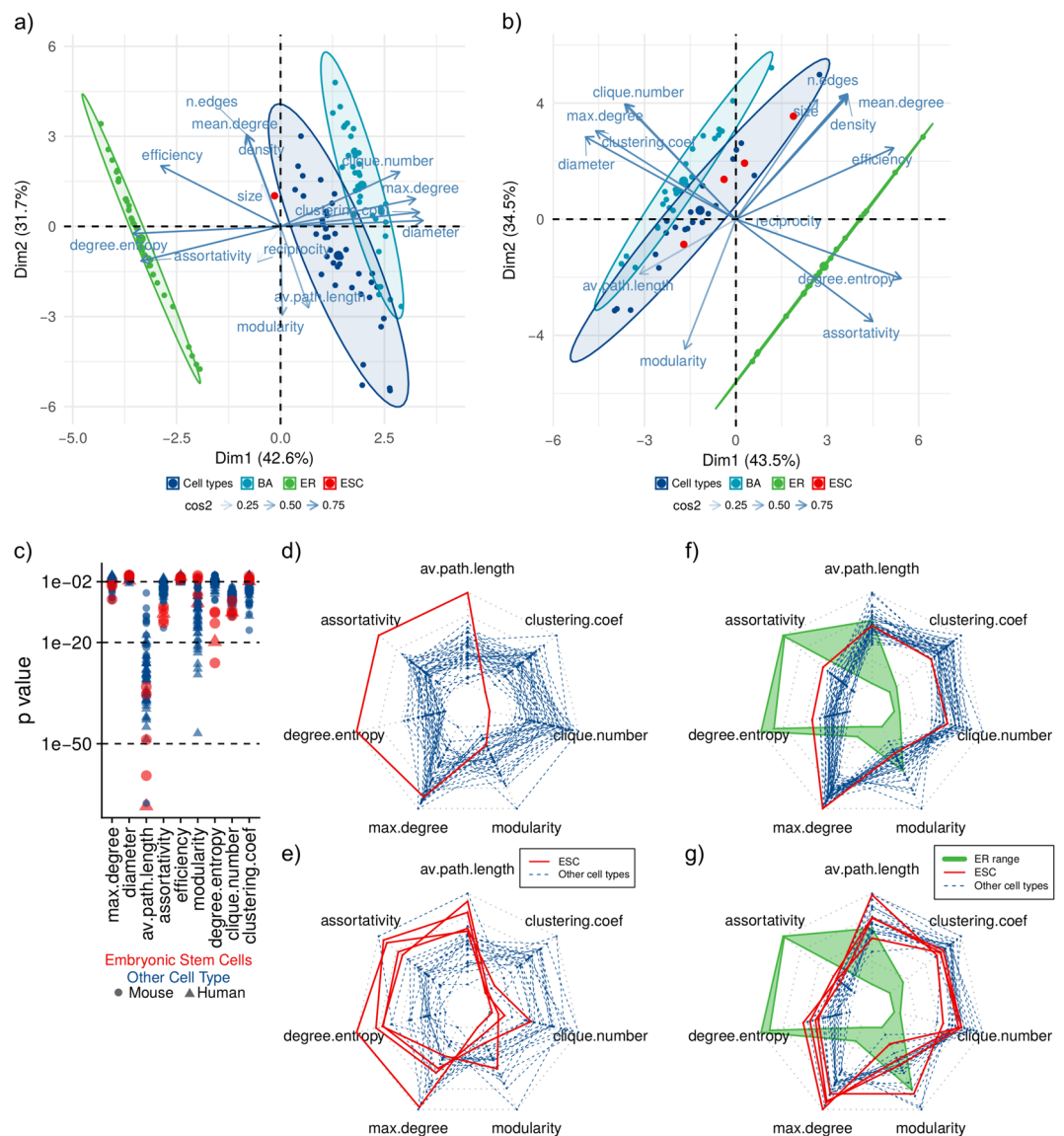


**Figure 4.** Networks comparison with ER and BA model networks. **(a)** Heatmap human and mouse cell types dissimilarity to model networks. **(b)** Scatterplot of cell types dissimilarity to model networks, dashed lines in both axis correspond to the 25, 50 and 75 percentiles of both measurements. **(c)** Distribution of  $D$  values for ESCs and other cell types in human and mouse, respectively. Distributions correspond to the dissimilarity with each of the 100 simulated model networks, black dots correspond to distribution median.

level, the distinctive network structure might capture an underlying, more permissive chromatin accessibility state, which has been previously hypothesized to underlie ESC behavior<sup>14,22</sup>. We tested this hypothesis empirically by directly analyzing DNase-seq chromatin accessibility data from the Roadmap Epigenomics project<sup>23</sup>, comparing samples corresponding to ESCs and adult differentiated cell types.

We compared accessibility signal (normalized counts) across all gene promoters, TF promoters only, and enhancers, considering these entities key regulatory elements in transcriptional networks (REs). Overall, REs display higher median accessibility in ESC than in adult samples (Fig. 7a-c). To test group differences between ESCs and adult tissues, we defined for each regulatory region a mean accessibility score, and found that REs are significantly more accessible in ESCs than adult tissues in the three cases (Fig. 7d-f).

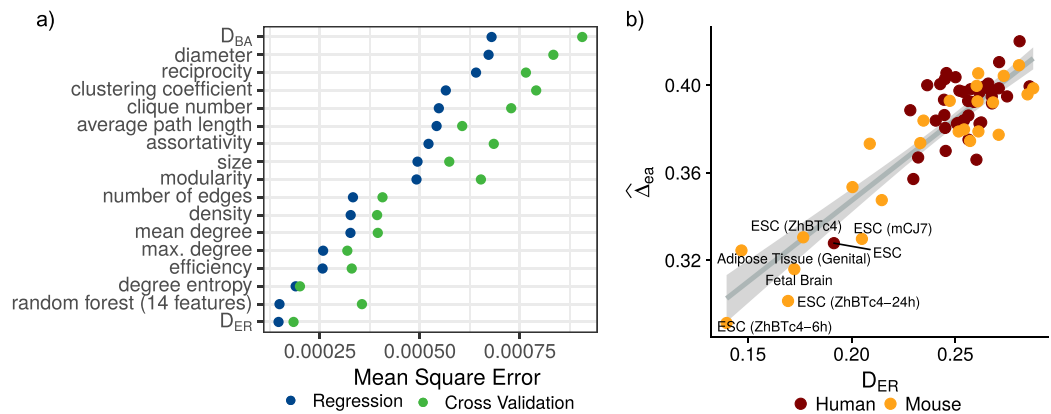




**Figure 5.** Comparison of network features with random model analogs. **(a)** Human and **(b)** mouse network features PCA including real, ER, and BA networks. **(c)** Network features p values comparing real features with BA model analogs. **(d)** Human and **(e)** mouse radar plots of network features z-score compared to BA model analogs. **(f)** Human and **(g)** mouse features measures, green polygons show ER networks' range for each feature. ESC values are shown in red solid lines, and the other cell types are shown in blue dashed lines.

The accessibility distribution reveals a higher median accessibility in regulatory regions within ESCs and shorter tails in the extremes of the distribution, relative to adult samples. This points to a more evenly distributed activity among REs, a pattern particularly pronounced in TF promoters. To quantify this observation, we measured the entropy of the accessibility distribution. TF promoters and enhancers show a significantly higher entropy in ESCs compared to adult differentiated cell types (Wilcoxon,  $p = < 0.023$ ) (Fig. 7h,i). This result indicates that the main elements of the regulatory circuits specifying cell-identity (enhancers and TFs)<sup>24</sup>, display a distinctive, promiscuous activity (as approximated by accessibility) in ESCs. The reduction of uncertainty in RE activity observed in adult differentiated cell types evidences a more restrictive epigenomic state, in which some TFs and enhancers have high activity and influence on the identity of the cell state. On the other hand, the state of uncertainty in the accessibility of the REs resulting in permissive global activity of TFs and target REs in ESCs may be ultimately manifested in the pluripotent, undecided, and promiscuous nature characteristic of ESCs<sup>22,25</sup>. These contrasting permissive and restrictive patterns of accessibility, in particular in the neighborhood of TF TSSs, is captured in the network structures analyzed herein.

The more accessible, permissive, and promiscuous activity of regulatory elements and regulators (TFs) in ESC populations is consistent with both their pluripotent nature and with an increased robustness of the TF networks characterizing their state.



**Figure 6.** Predictive models. (a) Mean square error for linear regressions of network vulnerability using each feature as predictor and random forest using 14 topological features. (b) Linear regression of network vulnerability predicted by networks' dissimilarity to Erdős-Rényi model network.

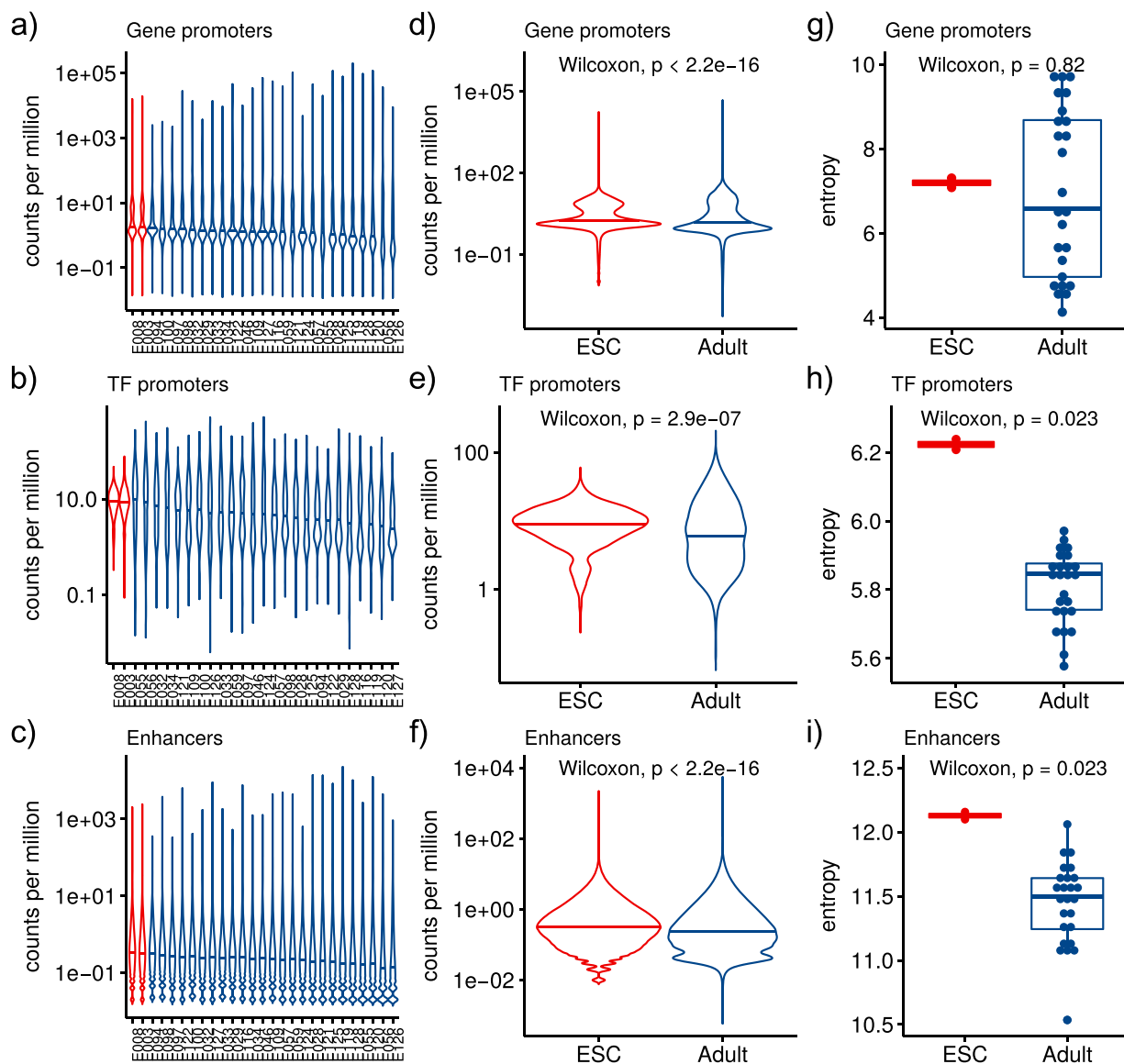
## Discussion

It has been pointed out that insights into the interplay between network structure and dynamics are needed in order to ultimately understand the cell's functional organization<sup>2</sup>. Here we studied TF networks' structure with the goal of better understanding the global behavior of different tissues. As a simple operational approximation, we represented the cell using tissue-specific TF networks. We frame the problem in terms of global structural robustness, a systemic behavior approximated by the vulnerability of networks to both random failure and directed perturbations<sup>2,18</sup>. We found that structural robustness varies significantly across tissues with different levels of differentiation. Interestingly, within the datasets analyzed in both human and mouse, the most robust tissue was also the least differentiated: embryonic stem cells.

Complex network theory has shown the coexistence of extremes in robustness and fragility ("robust yet fragile") in real-world networks, due to the widespread power-law connectivity distribution associated with complex networks<sup>18,26</sup>. The networks underlying ESCs are the most robust against random failure as well as the least fragile against directed attacks, somehow being able to negotiate the observed trade-off between robustness and fragility. It is known that deviation from the long-tail of theoretical networks with power-law degree distribution reduces the effectiveness of an attack strategy based on targeting the highly connected nodes<sup>18</sup>. Although all the TF networks analyzed here do have a long-tailed degree distribution, they deviate from theoretical power-law degree distributions (see Supplementary Figs 4 and 5). We analyzed this deviation from a canonical scale-free network by measuring each network's dissimilarity to theoretical model networks with homogeneous and scale-free topologies. This comparison further exposed the structural heterogeneity among tissues, and the deviating behavior of both undifferentiated (i.e., ESCs) and differentiated tissues. Furthermore, within the proposed analysis framework, the relative (dis)similarity between a target network and analogous theoretical networks provides insights into the topological characteristics underlying its robustness. For example, the higher structural robustness of ESC networks is explained by its closer topological resemblance to an Erdős-Rényi homogeneous random network, relative to differentiated cell types.

In terms of biological properties, our results suggest that ESC state might be able to withstand more and different kinds of errors, due to a more homogeneous network topology. This topological arrangement implies that its main regulator TFs act upon a less constrained chromatin landscape, allowing them to explore it more freely than in differentiated cell types. We further explored this idea by directly analyzing accessibility data at genome REs (TF promoters and enhancers), comparing ESCs and adult differentiated tissues. We found ESCs have a significantly higher accessibility at regulatory elements compared with differentiated tissues. ESCs also have a more evenly distributed accessibility among REs as shown by a higher entropy in its distribution (Fig. 7). Consistent with our results, several studies show that ESCs nuclear DNA is organized in an unusual way, in which chromatin appears to be more "open" than in differentiated cells<sup>27</sup>. Some of these findings are that histones and non-histones proteins are more loosely bound to DNA in ESC<sup>28</sup>, constitutive heterochromatin is more dispersed<sup>28,29</sup>, modifications associated with silent chromatin are depleted, while those associated with transcriptional activity are globally enriched<sup>28,29</sup>. These data has lead to consider stem and dedifferentiated cells as a state of loose regulation, differentiation being considered as a process of increasing chromatin repression<sup>27,28,30</sup>. Our results that show ESCs have a more homogeneous and structurally robust TF network topology can be considered a consequence of this loose regulation state in ESCs.

Previous studies have found a correlation between the level of uncertainty in the expression profile of a cell's signaling network and its differentiation potential (pluripotency)<sup>15,16</sup>. In other words, pluripotent cells can be characterized by a state of high uncertainty, where molecules from opposite lineages are promiscuously and simultaneously expressed. This uncertain state seems to mechanistically promote a cell-fate decision, due to its instability<sup>31,32</sup>. Entropy-based measures of uncertainty have been shown to capture such degree of instability and therefore pluripotency: lineage committed cells would have reduced entropy relative to progenitors, as differentiation is associated with the predominant expression of one of the mutually competing transcriptional programs. Consistent with this view, a network entropy measure integrating tissue-specific transcriptomic profiles with a protein interaction network, has effectively quantified cellular pluripotency using bulk<sup>15</sup> and single-cell data<sup>16</sup>.



**Figure 7.** Chromatin accessibility in ESCs and adult tissue samples from Roadmap Epigenomics data<sup>23</sup>. Number of DNase-seq tags per million reads in (a) gene promoters, (b) TF promoters, and (c) enhancers for each sample. Group mean accessibility distribution in (d) gene promoters, (e) TF promoters, and (f) enhancers. Boxplot of groups accessibility distribution entropy in (g) gene promoters, (h) TF promoters, and (i) enhancers. Horizontal lines inside violin plots correspond to the distribution median.

In the present study we found that the structural robustness of a transcription factor network clearly discriminate ESCs from differentiated cell types. Unlike transcriptomic analyses, however, this property does not seem to correlate with cellular differentiation potential within specific lineages. One potential interpretation for this observation is that the analyzed networks may highlight differences in chromatin organization that might anticipate transcriptional differences between cell types. On the other hand, the inability of these measures to distinguish between multipotent and fully differentiated cell types could stem from a lack of resolution to capture more subtle differences in network arrangement, or from the loss of information during TF networks inference due to the averaging intrinsic to bulk DNA-seq data. Nonetheless, our results do highlight an association between pluripotency and uncertainty of the regulatory network state, as measured by the entropy of chromatin accessibility profiles. This observation is consistent with the general model of a molecularly promiscuous cellular state underlying pluripotency. Here uncertainty is measured from chromatin accessibility profiles, while previous, higher resolution studies used transcriptomic data<sup>15,16</sup>. An interesting research direction would be to study the precise relation between the two measures of entropy, linking epigenomic structural data with transcriptomic profiles. In particular the recent development of single-cell resolution chromatin accessibility<sup>33</sup> and transcriptomic<sup>34</sup> profiling technologies might enable disentangling associations between multiple levels of regulation, perhaps overcoming the limitations of inferring TF networks based on bulk data alone.

It is well known that network topology plays a central role in dynamical behavior. In the cellular context, gene regulatory networks orchestrate cellular behavior<sup>35</sup>. Theoretical studies have previously analyzed the interplay between structure and dynamics using random Boolean networks<sup>36,37</sup>. Networks with a homogeneous topology and relatively high connectivity require fine tuned activation parameters in order to have a stable behavior, and to avoid chaotic dynamics<sup>36,37</sup>. This result seems inconsistent with the nature of real biological systems, which have a stable behavior despite fluctuations in surrounding environmental parameters. In other words, resilience is a characteristic of biological systems. Interestingly, for networks with a scale-free topology stable behavior emerges without the fine tuning requirement<sup>36,37</sup>. Considering our results in this structure/dynamics context, the higher homogeneity found in the ESC networks is likely to produce less ordered dynamics than more differentiated tissues, which, at the same time would allow them to explore more freely the state space and to reach multiple different network states. Interestingly, this view is consistent with the observed high heterogeneity in gene expression and with the balance between robustness and plasticity characteristic of ESCs<sup>15,25,38,39</sup>. Although we did not consider dynamical analysis in this study, but rather limited ourselves to the empirical, structural characterization of the networks and their behavior, disentangling structure and dynamics will be the focus of future work.

Summarizing, in light of the amount of data on biological interactions being generated in the post-genomic era, a systems level perspective is required to gain understanding of the biological systems as a whole. Our structural analysis of tissue specific TF networks aims at that objective, trying to find a connection between transcriptional networks structural heterogeneity and biological phenotypes. Our treatment of structural robustness as a network systems-level behavior revealed differences among cell types that could be dissected further through topological analyses and related to chromatin accessibility profile at REs. We want to stress the applicability of our comparison of real world complex networks not only for a structural characterization, but also as an approximation to their possible dynamic behaviors. Finally, the empirical analysis framework proposed here can be applied to any set of related networks whose structural heterogeneity is suspected to underly differential real life behavior.

## Methods

**Transcription Factor Networks.** Human and mouse transcription factor networks (TFNs) were constructed based on DNase-seq data and digital genomic footprinting as shown in<sup>9,14</sup>. Human networks set include 41 distinct cell and tissue specific networks composed of 493 to 533 sequence-specific transcription factors. Mouse networks set include 25 cell and tissue specific networks composed of 555 to 583 sequence-specific transcription factors. For simplicity, we use the term tissue-specific through the text to refer to both cell type and tissue. Network data were downloaded from <https://www.regulatorynetworks.org/>. Most current versions for human (v09162013) and mouse (v12032013) were used.

**Modeling topological robustness.** Topological robustness was approximated by profiling the network's behavior in response to random and directed structural perturbations. Site percolation was used as a process to model component failure using computer simulations<sup>6</sup>. Increasing fractions of a network's vertices were removed, along with the edges connected to those vertices. Following<sup>6,40</sup> a percolation process was considered in the general sense – i.e., including different ways of vertex removal. The error experiments performed correspond to the simplest percolation process where a fraction of vertices was chosen uniformly at random and removed. For every network, error experiments were repeated 1000 times and the mean error behavior was calculated. Directed (Attack) experiments were simulated by removing vertices in decreasing order of centrality based on vertex degree. Nodes were progressively removed from one to a hundred percent of nodes.

**Quantifying network structural robustness.** Two quantitative measures of network damage were used to characterize the phenomenology associated to the damage process applied to each TF network. As a first approximation, the macroscopic (systemic) behavior of the networks in response to damage was characterized by the evolution of the giant component size relative to its initial value as a function of the fraction of removed vertices  $f$  ( $S_f/S_0$ ). As an additional approximation, the global efficiency  $E$  of a network was used to quantify how communication becomes less efficient as damage increases, this measure was also calculated relative to its initial value and as a function of the fraction of removed vertices  $f$ . The latter measure assumes that the efficiency for sending information between two vertices  $i$  and  $j$  is proportional to the reciprocal of their distance, and is calculated as follows<sup>7,8</sup>:

$$E = \frac{1}{N(N-1)} \sum_{i \neq j} \frac{1}{d_{ij}} \quad (1)$$

The measure  $E$  corresponds to the average inverse geodesic length – i.e., the harmonic mean of the geodesic distances<sup>7</sup>:

$$h = \frac{1}{E} \quad (2)$$

**Error-Attack Deviation and vulnerability calculation.** The measure error-attack deviation  $\Delta_{ea}$  introduced herein, was used to quantify the degree of robustness to attacks relative to that against errors. The metric is simply the root mean square deviation between the observed error and the attack behaviors:

$$\Delta_{ea} = \sqrt{\frac{1}{n} \sum_f (e_f - a_f)^2} \quad (3)$$

where  $e_f(a_f)$  represents the a normalized measured of damage behavior under the random or (directed) removal of a fraction  $f$  of nodes. In this study  $S_f/S_0$  and  $E_f/E_0$  were used as damage measures (see Results).

We defined network vulnerability ( $\widehat{\Delta ea}$ ) as the mean between error-attack deviation to giant component size and efficiency:

$$\widehat{\Delta ea} = \frac{\Delta ea_{S_f/S_0} + \Delta ea_{E_f/E_0}}{2} \quad (4)$$

**Networks Topological Characterization.** Networks' topology was analyzed by quantifying topological dissimilarity and measuring 14 structural features commonly used in complex network theory<sup>6,7</sup>.

**Network dissimilarity.** Network dissimilarity measurement was done following the approach proposed by Shieber *et al.*<sup>19</sup>. This method compares networks topology based on quantifying differences among node distance probability distributions, representing all nodes connectivity distances, extracted from the networks. It returns non-zero values only for non-isomorphic graphs, and quantifies structural topological differences that have an impact on information flow through the network. We measured network dissimilarity following the algorithm proposed in<sup>19</sup>, using the suggested parameters.

**Networks structural characterization.** We described networks' topology by measuring 14 features: number of nodes, number of edges, mean degree, diameter, maximum degree, average path length, density, clustering coefficient, assortativity, efficiency, modularity, degree entropy, clique number, and reciprocity. Following the measurement definitions in<sup>7</sup>.

**Null models.** To compare cell type networks with random models, we generated random networks with the same number of nodes and links. Two sets of random networks were created: one set following Erdős-Rényi model (ER networks) with exponential degree distribution, and the second set following Barabási-Albert model of growing networks with power-law degree distribution (BA networks). In order for the BA networks to have an equivalent number of edges to its real counterpart, the number of outgoing edges added to each new node in the network was taken from the out degree distribution of the real network.

For each real network, 100 ER and BA random networks were created. Every random network was structurally characterized measuring the 14 topological features measured in the real networks, and dissimilarity to its real equivalent was quantified. Mean values for the dissimilarity and topological features were estimated for each ensemble of random networks.

**Features significance with respect to BA analogs.** For each cell type, we constructed a feature BA analog expected distribution from the feature's value in the 100 analog random BA networks. We then calculated the real feature z-score with respect to the BA expected distribution and using this z-score we obtained the p-value for each feature in every network.

**Predictive modeling.** Predictive models were fitted using networks' vulnerability as a response variable and structural features as predictors.

First we fitted a linear regression predicting  $\widehat{\Delta ea}$  using the 14 statistical features we measures, plus the network's dissimilarity to its ER analogs ( $D_{ER}$ ) and to its BA analogs ( $D_{BA}$ ) as predictors. The second model we fitted was a random forest regression, predicting  $\widehat{\Delta ea}$  from the 14 topological features measured above, this model was created with 1000 trees. Features' influence on the random forest model was measured by the mean decrease in mean square error. As a way to evaluate the models' accuracy, we performed a five-fold cross validation of both models, keeping the test mean square error as accuracy measurement.

**Comparing DNase-seq data chromatin accessibility.** DNase-seq alignment files were downloaded from the Roadmap Epigenomics data portal at [https://egg2.wustl.edu/roadmap/web\\_portal/processed\\_data.html](https://egg2.wustl.edu/roadmap/web_portal/processed_data.html)<sup>23</sup>. Only samples corresponding to ESC and Adult anatomical groups were kept. Aligned reads were mapped to promoters, and enhancers. Gene promoters were defined as 5 kb regions surrounding the TSS from Genecode database [www.genecodegenes.org/releases/current.html](http://www.genecodegenes.org/releases/current.html), from these gene promoters we extracted 600 TFs present at HOCOMOCO database <https://autosome.ru/hocomoco/><sup>41</sup> to define the TF promoters. Enhancers regions were defined based on Roadmap ChromHMM segmentations data, considering only the distal, non-genic enhancer state from the 15-state model. Reads mapping target regions were aggregated using bedops with the bedmap command<sup>42</sup>. A group mean accessibility score was defined among all ESC and adult samples in every genomic region by calculating mean accessibility across samples of the same group.

**Implementation.** All the methods presented here were implemented using the R statistical programming environment [www.R-project.org](http://www.R-project.org) and the igraph package<sup>43</sup>.

## References

- Huang, S. Back to the biology in systems biology: What can we learn from biomolecular networks? *Briefings in functional genomics & proteomics* **2**, 279–297 (2004).
- Barabasi, A.-L. & Oltvai, Z. N. Network biology: understanding the cell's functional organization. *Nature reviews genetics* **5**, 101–113 (2004).
- Babu, M. M., Teichmann, S. A. & Aravind, L. Evolutionary dynamics of prokaryotic transcriptional regulatory networks. *Journal of molecular biology* **358**, 614–633 (2006).
- Thiele, I. *et al.* A community-driven global reconstruction of human metabolism. *Nature biotechnology* **31**, 419–425 (2013).

5. Li, T. *et al.* A scored human protein-protein interaction network to catalyze genomic interpretation. *Nature methods* (2016).
6. Newman, M. *Networks: an introduction* (OUP Oxford, 2010).
7. Costa, Ld. F., Rodrigues, F. A., Traverso, G. & Villas Boas, P. R. Characterization of complex networks: A survey of measurements. *Advances in Physics* **56**, 167–242 (2007).
8. Barrat, A., Barthélemy, M. & Vespignani, A. *Dynamical processes on complex networks* (Cambridge University Press, 2008).
9. Neph, S. *et al.* Circuitry and dynamics of human transcription factor regulatory networks. *Cell* **150**, 1274–1286 (2012).
10. Marbach, D. *et al.* Tissue-specific regulatory circuits reveal variable modular perturbations across complex diseases. *Nature methods* (2016).
11. Hesselberth, J. R. *et al.* Global mapping of protein-DNA interactions *in vivo* by digital genomic footprinting. *Nature Methods* **6**, 283–289, <https://doi.org/10.1038/nmeth.1313> (2009).
12. Sullivan, A. M., Bubb, K. L., Sandstrom, R., Stamatoyannopoulos, J. A. & Queitsch, C. Dnase i hypersensitivity mapping, genomic footprinting, and transcription factor networks in plants. *Current Plant Biology* **3**, 40–47 (2015).
13. Vierstra, J. & Stamatoyannopoulos, J. A. Genomic footprinting. *Nature methods* **13**, 213–221 (2016).
14. Stergachis, A. B. *et al.* Conservation of trans-acting circuitry during mammalian regulatory evolution. *Nature* **515**, 365–370 (2014).
15. Banerji, C. R. S. *et al.* Cellular network entropy as the energy potential in waddington's differentiation landscape. *Scientific Reports* **3**, 3039 (2013).
16. Teschendorff, A. E. & Enver, T. Single-cell entropy for accurate estimation of differentiation potency from a cell's transcriptome. *Nature Communications* **8**, 15599, <https://doi.org/10.1038/ncomms15599> (2017).
17. Kolaczyk, E. D. & Csárdi, G. *Statistical analysis of network data with R*, vol. 65 (Springer, 2014).
18. Albert, R., Jeong, H. & Barabási, A.-L. Error and attack tolerance of complex networks. *nature* **406**, 378–382 (2000).
19. Schieber, T. A. *et al.* Quantification of network structural dissimilarities. *Nature Communications* **8**, 13928, <https://doi.org/10.1038/ncomms13928> (2017).
20. Zhang, Z. & Zhang, J. A Big World Inside Small-World Networks. *PLoS One* **4**, e5686, <https://doi.org/10.1371/journal.pone.0005686> (2009).
21. Barabási, A.-L. & Albert, R. Emergence of scaling in random networks. *Science* **286**, 509–512 (1999).
22. MacArthur, B. D. & Lemischka, I. R. Statistical mechanics of pluripotency. *Cell* **154**, 484–489 (2013).
23. Consortium, R. E. *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).
24. Maston, G. A., Evans, S. K. & Green, M. R. Transcriptional regulatory elements in the human genome. *Annual Review of Genomics and Human Genetics* **7**, 29–59 (2006).
25. Garcia-Ojalvo, J., Arias, A. M. & Martinez Arias, A. Towards a statistical mechanics of cell fate decisions. *Current Opinion in Genetics and Development* **22**, 619–626, <https://www.ncbi.nlm.nih.gov/pubmed/23200114> (2012).
26. Doyle, J. *et al.* The “robust yet fragile” nature of the internet. *Proceedings of the National Academy of Sciences USA* **102**, 14479–14502 (2005).
27. Turner, B. M. Open Chromatin and Hypertranscription in Embryonic Stem Cells. *Cell Stem Cell* **2**, 408–410 (2008).
28. Meshorer, E. & Misteli, T. Chromatin in pluripotent embryonic stem cells and differentiation. *Nature Reviews Molecular Cell Biology* **7**, 540–546 (2006).
29. Spivakov, M. & Fisher, A. G. Epigenetic signatures of stem-cell identity. *Nature reviews. Genetics* **8**, 263–271 (2007).
30. Marks, H. *et al.* The transcriptional and epigenomic foundations of ground state pluripotency. *Cell* **149**, 590–604 (2012).
31. Huang, S., Guo, Y.-P., May, G. & Enver, T. Bifurcation dynamics in lineage-commitment in bipotent progenitor cells. *Developmental biology* **305**, 695–713 (2007).
32. Zhou, J. X. & Huang, S. Understanding gene circuits at cell-fate branch points for rational cell reprogramming. *Trends in Genetics* **27**, 55–62 (2011).
33. Buenrostro, J. D. *et al.* Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* **523**, 486 (2015).
34. Marr, C., Zhou, J. X. & Huang, S. Single-cell gene expression profiling and cell state dynamics: collecting data, correlating data points and connecting the dots. *Current opinion in biotechnology* **39**, 207–214 (2016).
35. Davila-Velderrain, J., Martinez-Garcia, J. C. & Alvarez-Buylla, E. R. Modeling the epigenetic attractors landscape: toward a post-genomic mechanistic understanding of development. *Frontiers in genetics* **6** (2015).
36. Aldana, M. Boolean dynamics of networks with scale-free topology. *Physica D: Nonlinear Phenomena* **185**, 45–66 (2003).
37. Valverde, S., Ohse, S., Turalska, M., West, B. J. & Garcia-Ojalvo, J. Structural determinants of criticality in biological networks. *Frontiers in physiology* **6**, 127, <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4424853&tool=pmcentrez&rendertype=abstract> (2015).
38. Huang, S. Systems biology of stem cells: three useful perspectives to help overcome the paradigm of linear pathways. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* **2247–2259**, <https://doi.org/10.1098/rstb.2011.0008>.
39. Kaneko, K. Characterization of stem cells and cancer cells on the basis of gene expression profile stability, plasticity, and robustness: Dynamical systems theory of gene expressions under cell-cell interaction explains mutational robustness of differentiated cells. *BioEssays* **33**, 403–413, <https://www.ncbi.nlm.nih.gov/pubmed/21538414> (2011).
40. Callaway, D. S., Newman, M. E., Strogatz, S. H. & Watts, D. J. Network robustness and fragility: Percolation on random graphs. *Physical review letters* **85**, 5468 (2000).
41. Kulakovskiy, I. V. *et al.* HOCOMOCO: Expansion and enhancement of the collection of transcription factor binding sites models. *Nucleic Acids Research* **44**, D116–D125 (2016).
42. Neph, S. *et al.* Bedops: high-performance genomic feature operations. *Bioinformatics* **28**, 1919–1920, <https://doi.org/10.1093/bioinformatics/bts277> (2012).
43. Csardi, G. & Nepusz, T. The igraph software package for complex network research. *InterJournal, Complex Systems* **1695**, 1–9 (2006).

## Acknowledgements

This work was supported by Consejo Nacional de Ciencia y Tecnología, (CONACYT: 240180, 180380, 2015-01-687) and UNAM-DGAPA-PAPIIT (IN211516, ININ208517, IN205517, IN204217). J.C.P. is a doctoral student from Programa de Doctorado en Ciencias Biomédicas, Universidad Nacional Autónoma de México (UNAM) and received fellowship 446988 from CONACYT. We thank Diana Romo for logistic support.

## Author Contributions

J.D.V. designed and coordinated the study. J.C.P. conducted the analyses. J.C.P. and J.D.V. analyzed the results, and wrote the manuscript. E.A.B. provided resources and discussed the problem and results. All authors read and approved the manuscript.

## Additional Information

**Supplementary information** accompanies this paper at <https://doi.org/10.1038/s41598-018-32020-1>.

**Competing Interests:** The authors declare no competing interests.

**Publisher's note:** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2018

# D In vivo and in vitro human gene essentiality estimations capture constrasting functional constraints

Artículo publicado en 2021 en *Nucleic Acids Research Genomics and Bioinformatics*.



# ***In vivo* and *in vitro* human gene essentiality estimations capture contrasting functional constraints**

Jose Luis Caldu-Primo<sup>1,2</sup>, Jorge Armando Verduzco-Martínez<sup>3</sup>,  
Elena R. Alvarez-Buylla<sup>1,2,\*</sup> and Jose Davila-Velderrain<sup>4,5,\*</sup>

<sup>1</sup>Instituto de Ecología, Universidad Nacional Autónoma de México, Cd. Universitaria, CDMX., 04510, México, <sup>2</sup>Centro de Ciencias de la Complejidad (C3), Universidad Nacional Autónoma de México, Cd. Universitaria, CDMX., 04510, México, <sup>3</sup>Departamento de Biología Celular y Genética, Facultad de Ciencias Biológicas, Universidad Autónoma de Nuevo León, San Nicolás de los Garza, Nuevo León, 66400, México, <sup>4</sup>MIT Computer Science and Artificial Intelligence Laboratory, Cambridge, MA 02139, USA and <sup>5</sup>Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA

Received May 06, 2021; Revised June 18, 2021; Editorial Decision June 21, 2021; Accepted July 06, 2021

## **ABSTRACT**

**Gene essentiality estimation is a popular empirical approach to link genotypes to phenotypes. In humans, essentiality is estimated based on loss-of-function (LoF) mutation intolerance, either from population exome sequencing (*in vivo*) data or CRISPR-based *in vitro* perturbation experiments. Both approaches identify genes presumed to have detrimental consequences on the organism upon mutation. Are these genes constrained by having key cellular/organismal roles? Do *in vivo* and *in vitro* estimations equally recover these constraints? Insights into these questions have important implications in generalizing observations from cell models and interpreting disease risk genes. To empirically address these questions, we integrate genome-scale datasets and compare structural, functional and evolutionary features of essential genes versus genes with extremely high mutational tolerance. We found that essentiality estimates do recover functional constraints. However, the organismal or cellular context of estimation leads to functionally contrasting properties underlying the constraint. Our results suggest that depletion of LoF mutations in human populations effectively captures organismal-level functional constraints not experimentally accessible through CRISPR-based screens. Finally, we identify a set of genes (*OrgEssential*), which are mutationally intolerant *in vivo* but highly tolerant *in vitro*. These genes drive observed functional constraint differences and have an unexpected preference for nervous system expression.**

## **INTRODUCTION**

Understanding the patterns and phenotypic consequences of genetic alterations is a fundamental problem in evolution and development (1–4). A popular empirical approach to link genotypes to phenotypes is by estimating the degree of *essentiality* of a gene. A gene is considered ‘essential’ if it is required to sustain life in cells or whole organisms, and this requirement is often estimated by experimental perturbations (5,6). The study of essential genes was originally conducted on prokaryotes, due to their accessibility to genetic manipulation. More recently, however, gene essentiality has been estimated in multicellular eukaryotes, including mammals (7). Despite the absolute character of the ‘essential’ gene denomination, data from multiple studies in model organisms have shown strong context dependency: genes are required or not for survival depending on environmental conditions and developmental stages (5,8,9).

The advent of sequencing technologies and gene editing techniques enabled the estimation of gene essentiality in humans (6). The problem has been addressed following two approaches. On one hand, systematic testing of gene silencing effects on human cell cultures identifies genes that affect cell viability or optimal fitness upon perturbation (10–13). On the other hand, population-level statistical estimates of unexpected mutational depletion identifies genes presumed to be subjected to functional constraints (14). Both approaches aim at ranking genes according to their effect on the organism (or cell) upon loss-of-function mutations. However, given the context dependency of gene essentiality, and the differences in the organizational level at which the effects of genotypic changes are assessed, the parallels of the two types of essentiality approximations are unclear.

*In vitro* screens of mutation tolerance identify genes with an immediate effect on cell proliferation and viabil-

\*To whom correspondence should be addressed. Tel: +1 617 253 3434; Email: [jdavilav@mit.edu](mailto:jdavilav@mit.edu)  
Correspondence may also be addressed to Elena R. Alvarez-Buylla. Tel: +1 617 253 3434; Email: [elenabuylla@protonmail.com](mailto:elenabuylla@protonmail.com)

ity; consequently, the corresponding essentiality estimates depend on the specific cell line and culture conditions being tested. Furthermore, cell culture experiments do not capture developmental and functional constraints intrinsic to the organism. *In vitro* estimation of gene essentiality is thus inevitably tailored to cell viability. On the other hand, ‘*in vivo*’ measures of mutational tolerance estimated from population-level genetic variation score genes according to the prevalence/depletion of loss-of-function (LoF) mutations. Genes under mutational constraint are assumed to be consistent with a scenario where purifying selection filters out protein-altering mutations with detrimental effects, thus eluding fixation within the population. In this sense, *in vivo* estimates of mutational tolerance are considered a proxy for the effect of mutations on organismal fitness. Such effect, in turn, mirrors to some extent the notion of essentiality in the context of population dynamics (5). Both estimation types (*in vitro* and *in vivo*) have been discussed within the context of human gene essentiality, nonetheless (6). Hereafter we use the terms cellular viability (CV) and organismal fitness (OF) to refer to the context in which human gene essentiality is estimated: by means of *in vitro* perturbation experiments (CV), or *in vivo* population-based mutation tolerance estimates (OF).

Notably, in both the CV and the OF context, a subset of mutational intolerant genes has been identified, leading to the idea of defining an ‘essential genome’ containing genes that do not tolerate mutations, and a ‘dispensable genome’ including mutation-tolerant genes (6,14). Intolerant genes (essential) are commonly of interest due to their potential detrimental effect on phenotype and disease association; however, highly tolerant genes (nonessential) might be relevant for evolvability, due to the plasticity they confer to the system at longer time-scales—for example, as sources of cryptic genetic variation (4) or possible editable links that integrate subsystems (15). Hereafter we will use the terms *tolerant* and *intolerant* to refer to human nonessential or essential genes as estimated by the degree of LoF mutation tolerance.

Despite the potential functional relevance of tolerant and intolerant genes, an understanding of the molecular determinants that discriminate between the two groups has been only partially explored for humans (6). Moreover, an understanding of the dependency of molecular determinants of gene essentiality on the differences between the operational context of estimation (CV, *in vitro* versus OF, *in vivo*) is lacking. To address these problems, here we systematically defined groups of human tolerant and intolerant genes and performed an integrative and comparative analysis of the structural, functional, and evolutionary features associated with gene essentiality. We analyzed the particularities and commonalities between genes that show extreme (in)tolerance to LoF mutation in a given context: CV, OF or both (Figure 1).

## MATERIALS AND METHODS

### Gene essentiality

Human gene essentiality estimations based on measures of tolerance to LoF mutations were taken from (6). Estimates

include the following scores based on the Exome Aggregation Consortium (ExAC) sample of 60 706 human exomes (14): residual variation intolerance score (RVIS) (16), EvoTol (17), missense Z-score (18), LoFtool (19), probability of haploinsufficiency (Phi) (20), probability of loss-of-function intolerance (pLI) (14) and selection coefficient against heterozygous loss-of-function (shet) (21). Scores based on cell culture perturbation-based experiments include data from KBM7, Raji, Jiyoye, HCT116 and K562 cell lines (12); the KBM7 cell line (10), and RPE1, GBM514, HeLa and DLD1 cell lines (22).

### Intrinsic structural disorder

Disorder predictions for each protein in the human proteome were generated at residue resolution using IUPred (23). A gene intrinsic disorder score was calculated by averaging the predicted residue scores over the corresponding protein. Scores range from 0 to 1, with higher scores indicating a higher propensity toward intrinsic disorder.

### Haploinsufficiency

A predictive genome-wide haploinsufficiency score (GHIS) was obtained from (24).

### Gene expression specificity

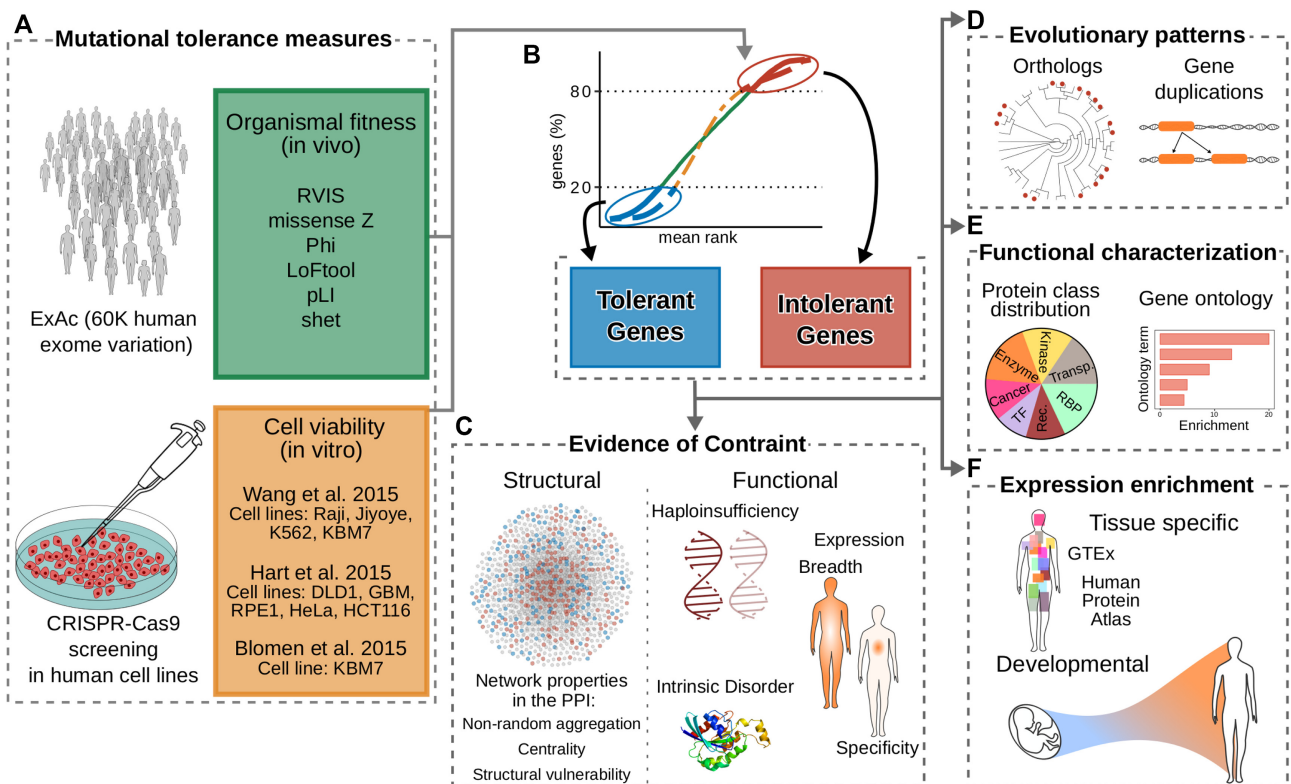
Reference RNA-seq data for human tissues was downloaded from the Genotype-Tissue Expression project (GTEx.v7) (25) (<https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-5214/>), and the Human protein atlas (HPA) (26) (<https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-2836/>). The GTEx dataset includes 53 tissues profiled from 961 donors. The HPA dataset includes 32 tissues profiled from 122 control subjects. For both datasets the median expression over replicates was considered as the expression value of the tissue. Expression breadth values for each gene were calculated as the fraction of tissues in which the gene is expressed, using an arbitrary cut-off value of 2 RPKM to determine expression. Expression specificity was measured using the Tau statistic on the same tissue-median matrix as computed in (27). Briefly, the Tau statistic is calculated as follows:

$$\tau = \frac{\sum_{i=1}^n (1 - \hat{x}_i)}{n - 1}; \hat{x}_i = \frac{x_i}{\max_{1 \leq i \leq n} (x_i)},$$

where  $x_i$  is the expression of gene  $x$  in tissue  $i$  and  $n$  is the number of tissues. From this definition, Tau varies from 0 to 1, with ubiquitously expressed genes having 0 value and extremely specific genes a value of 1.

### Protein classifications

Proteins were classified as transcription factors (TF), transporters, receptors, enzymes, peptidase, kinase, cancer-related, and RNA-binding proteins (RBP) based on combined curated annotations extracted from the Human Protein Atlas (<https://www.proteinatlas.org/humanproteome/proteinclasses>) (26), TF reference in (28), RBPs reference from (29), and transporters and receptors reported in (30).



**Figure 1.** Overview. (A) Mutational tolerance scores used to categorize human (in)tolerant genes. (B) Consensus mutational tolerance score derivation (mean rank distribution) and corresponding (in)tolerant gene sets. (C–F) Features considered as potential determinants of mutational constraint and gene essentiality, including structural and functional features (C), evolutionary (D), protein functional characterization (E) and expression enrichment (F).

### Evolutionary conservation

Comprehensive gene homology information for each human gene with respect to 187 species was extracted from Ensembl comparative genomics resources (31). Only one-to-one orthology relationships were considered to build a binary gene-species matrix. A gene conservation index was calculated for each human gene as the fraction of species having a corresponding ortholog (31). Gene duplication data were extracted from (32,33).

### Developmental annotations

Developmental expression classes and developmental process gene annotations were downloaded from the Online Gene Essentiality database (OGEE.v2) at (<http://ogee.medgenius.info/downloads/>) (33).

### Mutational tolerance gene group definition

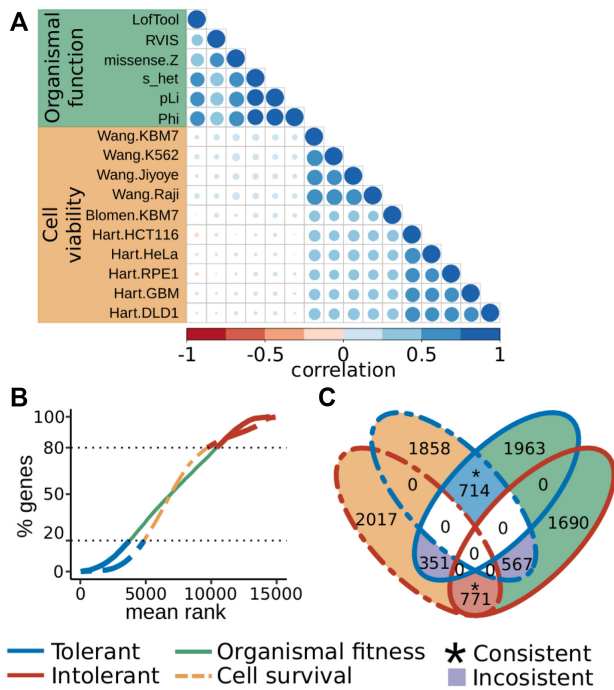
Mutational tolerance groups were defined based on consensus tolerance scores estimated by averaging gene ranks across available tolerance measures (OF,  $n = 6$ ; and CV,  $n = 10$  measures). For all measures, as reported in (6), values increase with the degree of intolerance to mutation: intolerant genes have high values. Tolerant genes were defined as the bottom 20% genes in the consensus score rank (lowest constraint) and Intolerant genes as the top 20% (highest constraint). The choice of cut-off values captures extreme

values from a long-tailed distribution, which approximates the cut-off proposed in (14) to define the widely used LoF-intolerance metric pLI (cut-off pLI > 0.9). Four additional subgroups were defined based on the patterns of overlap between intolerant and intolerant groups (Figure 2C): consistent tolerant genes ( $n = 714$  genes classified as tolerant in both conditions), consistent intolerant genes ( $n = 771$  genes classified as intolerant in both conditions), organismal intolerant but cellular tolerant genes (OI-CT) ( $n = 567$  genes classified as OF intolerant and CV tolerant), and cellular intolerant but organismal tolerant genes (CI-OT) ( $n = 351$  genes classified as OF tolerant and CV intolerant).

### Analysis of gene set aggregation and centrality in the PPI network

A reference human protein-protein interaction (PPI) network was obtained from (34). Briefly, the network is based on experimentally supported protein-protein interactions from different sources that through a stringent orthology mapping scheme recover 625 641 interactions among 17 530 human proteins.

The degree to which a set of genes is aggregated forming a neighborhood within the PPI network was quantified using three complementary approaches: (i) estimating the deviation of the size of the subnetwork produced by genes within the set and their interactions (module size) from expectation, (ii) estimating the strength of association among genes



**Figure 2.** Mutational tolerant and intolerant groups definition. (A) Correlation plot of mutational tolerance measures, adapted from (6). (B) Mutational tolerance measures mean rank distribution. Tolerant and intolerant gene sets are defined as the bottom and top 20%, respectively. (C) Venn diagram of the defined gene sets. Group intersections are highlighted to represent tolerant (714 genes) and intolerant (771) genes found consistently in both OF and CV groups, and inconsistent genes found in contradictory groups depending on the context.

within the set by clustering enrichment and (iii) contrasting observed pairwise gene network distances with expectation. Subgraph module size was calculated by counting the number of nodes ( $S_c$ ) and edges ( $C_c$ ) of the largest connected subgraph formed by proteins belonging to a given gene set. Clustering enrichment was measured using spatial analysis of network association, as implemented in SANTA (35). Pairwise shortest distance between every protein pair was measured using the *igraph* R package (36). The distance distribution of each gene set was characterized by calculating its minimum ( $D_s$ ) and mean ( $D_{sm}$ ) distances. Network centrality was calculated using three complementary measures: degree, betweenness and coreness. These measures were quantified for every node using the *igraph* R package (36). For each gene set and aggregation statistic, enrichment was calculated by estimating the deviation of the observed gene set average from that expected in a distribution obtained from 10 000 randomly sampled gene sets of the same size. Deviation was quantified with a  $z$ -score.

#### Network structural robustness analysis

Network robustness was characterized by measuring the effect on network structure of targeted removal of nodes according to mutational tolerance ranking and estimating its deviation from random expectation. Network structural response was assessed by calculating the number of nodes ( $S_f$ )

and edges ( $C_f$ ) in the perturbed largest connected component after removing a fraction ( $f$ ) of nodes relative to the unperturbed measures. For each measure, random expectation was estimated by removing fractions from 0.01 to 0.99 of randomly selected nodes 10 000 times.

#### Functional enrichment and protein class distribution

Over-representation of gene functional features (GHIS, ID, expression specificity, expression breadth, earliest stage expression and developmental process annotation) was estimated by contrasting the gene set average and the random expectation obtained from measuring the given feature in 10 000 randomly sampled same-sized gene sets. Deviation in protein class distribution among gene sets was assessed by quantifying the deviation of the percentage of genes belonging to each protein class from its random expectation as estimated from 10 000 randomly sampled same-sized gene sets.

#### Gene set evolutionary analysis

Starting from a binary gene-species matrix based on one-to-one orthology relationships, species were classified by taxonomic group resulting in: Archaea (21 species), Bacteria (99 species), Protozoa (16 species), Fungi (8 species), Plants (9 species), Invertebrates (24 species) and Vertebrates (10 species). Percentage of orthologs was calculated for each taxonomic group and mutational tolerance gene group. Orthology relationships were extracted from homologies reported in Ensembl Compara v101 (37).

#### Tissue specific and expression enrichment

Patterns of preferential expression across tissues were assessed using transcriptomic data (RNA-seq) from the Genotype-Tissue Expression Project (GTEx) (51 tissues) (25) and from the Human Protein Atlas (HPA) (32 tissues) (26). Gene expression enrichment was estimated by pairwise differential expression analysis across all tissue pairs. Differential expression was calculated using *voom* and *limFit* functions from the *limma* package in R (38). Expression enrichment scores are defined as the sum of the differential expression coefficients across comparisons, discarding genes with a Bonferroni corrected  $P$ -value  $>0.05$ . Tissue specificity of mutational tolerance groups was estimated by quantifying per tissue and gene groups the deviation from random expectation of expression enrichment scores. Random expectation was estimated by randomly sampling same-sized gene sets 10 000 times.

#### Gene set enrichment analysis

Gene ontology enrichment analysis was performed using DAVID (<https://david.ncifcrf.gov/summary.jsp>) (39,40).

#### GWAS trait associated genes enrichment

Genome-wide association studies (GWAS) data were downloaded from the NHGRI-EBI Catalog of human

genome-wide association studies (41). Significant associations and mapped genes as reported by the Catalog were for the following traits were considered: cancer (EFO\_1000654), type 2 diabetes (T2D, EFO\_0001360), Alzheimer disease (AD, EFO\_1001870), amyotrophic lateral sclerosis (ALS, EFO\_0001357), Parkinson's disease (PD, EFO\_0002508), schizophrenia (EFO\_0004609), major depression (EFO\_0009854), cognition (EFO\_0005229), intelligence (EFO\_0004337) and bipolar disorder (BP, EFO\_0009963). Over-representation of trait associated genes among the (in)consistency mutational tolerance classes was measured using Fisher's exact test as implemented in the R package SuperExactTest (42). Gene set enrichment analysis was performed using fgsea package (43) using GWAS associated genes, mutational tolerance mean ranking scores for OF and CV measures, and the rank difference between OF and CV measures as input. A functional classification of the top/bottom 15 genes in the rank difference distribution associated with any trait was defined using information from Uniprot (44).

#### Code and data availability

Code and data to reproduce results and all figures are available through GitHub: <https://github.com/jlcaldu/Gene-essentiality-analysis>.

## RESULTS

### Context-dependent mutational tolerance categorises human genes

To define genes with extreme (in)tolerance to detrimental mutation, as estimated from patterns of mutational depletion in exome sequencing data (OF) or fitness effects in CRISPR-based cell culture perturbation experiments (CV), we first calculated for each gene and context a consensus tolerance score by averaging gene ranks across available tolerance measures (OF,  $n = 6$ ; and CV,  $n = 10$  measures). The high pairwise correlation (average Pearson correlation = 0.59, 0.42; OF, CV) between individual measures within each context justifies the use of the proposed consensus score (Figure 2A). We then defined a set of mutation-intolerant (tolerant) genes based on the distribution of consensus tolerance scores. We used the 80th percentile of the distribution as arbitrary cut-off value, a choice that captures the extreme values observed in the long tailed distribution of the measures, and which approximates the cut-off proposed in (14) to define the widely used LoF-intolerance metric pLI (cut-off pLI > 0.9). In addition, we defined a contrasting, similar-sized set of mutation-tolerant (intolerant) genes by selecting the 20% bottom ranked genes of the consensus score distribution (Figure 2B). The size of the resulting gene sets are 3028 genes for both intolerant and tolerant OF groups and 3139 genes for both intolerant and tolerant CV groups. Exploring the intersection between gene sets, we identified 714 tolerant and 771 intolerant genes with consistent tolerance behavior across contexts. In contrast, we identified 918 inconsistent genes, whose tolerance behavior depends on the context (OF/CV) (Figure 2C).

### Structural and functional constraints predict mutational tolerance classes

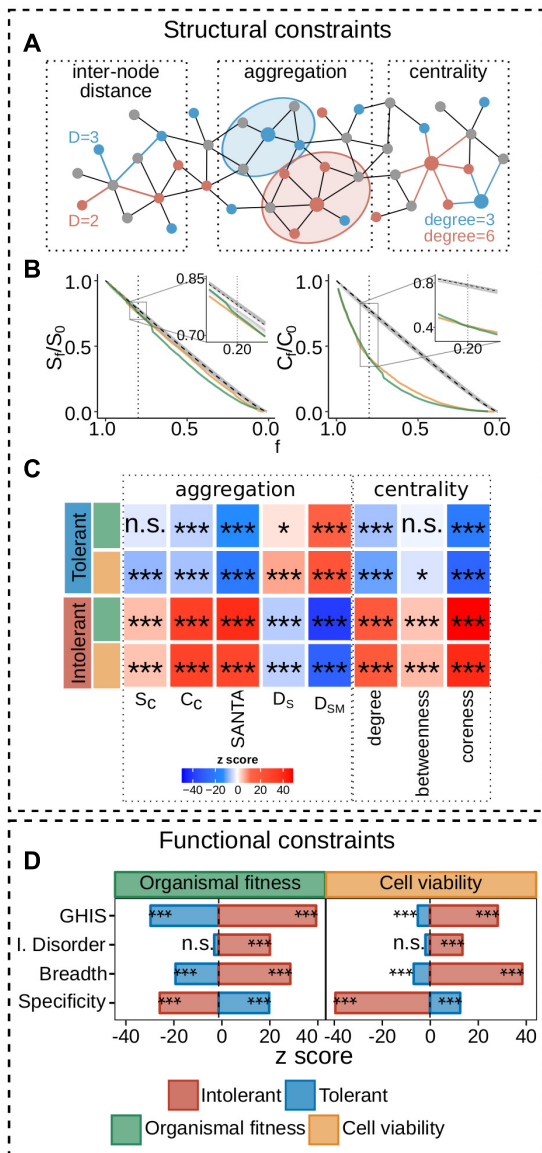
We next tested whether the different gene groups are distinctively associated with network structural, and with functional molecular properties. Following previous studies that point to a central role of essential genes in the interactome (45–48), we first asked if (in)tolerant genes have contrasting positions in the interactome and whether such pattern is consistent in genes affecting both organismal and cellular fitness. From a network perspective, in the context of the present study, we hypothesized that (i) mutational tolerance estimation may allow to identify evidence of a core constrained neighborhood within the human interactome, which is separated from a more peripheral, scattered layer formed by mutationally tolerant genes and (ii) given the central role of the intolerant neighborhood, perturbations affecting the corresponding genes are more likely to confer structural fragility to the entire system.

By measuring network features associated with node centrality and aggregation (Figure 3A, see Materials and Methods section), we confirmed that irrespective of the context of estimation, intolerant genes are aggregated and central in the interactome ( $P$  values < 0.001, two-sided  $z$  test); while tolerant genes consistently show the opposite behavior: loose aggregation and peripheral positioning (Figure 3C and Supplementary Figure S1). To test the vulnerability of the interactome to perturbations targeting tolerant or intolerant genes, we analyzed the network's behavior as a function of the progressive removal of nodes in decreasing order of mutational intolerance score (Materials and Methods section). This analysis further confirmed that there is a strong association between mutational patterns and the global structural properties of the interactome, revealing that intolerant gene removal produces a higher structural damage than random node removal (Figure 3B).

Our results suggest that global structural properties of the interactome are related to mutational tolerance classes. We reasoned that molecular properties suggestive of functional constraint might discriminate tolerance groups as well. By analyzing the degree of estimated haploinsufficiency (GHIS), protein intrinsic disorder (ID), expression breadth and specificity, we similarly found contrasting patterns between tolerant and intolerant genes. Intolerant genes are more likely to be haploinsufficient, to have more intrinsic disorder in protein structure, and to be more broadly expressed across tissues; while tolerant genes show exact opposite behavior (Figure 3D and Supplementary Figure S1). Together, our results confirm that mutationally tolerant and intolerant genes can be consistently discriminated by features indicative of structural and functional constraints, and that this property is independent of the context in which tolerance is estimated (OF or CV).

### Evolutionary history of tolerance gene classes

Gene essentiality and interactome centrality have been previously related to evolutionary conservation, with a tendency for topologically central and essential genes to be conserved (evolutionarily old) (49,50). We analyzed whether the tolerance gene groups identified here similarly have contrasting evolutionary conservation patterns, and



**Figure 3.** Structural and functional constraints. (A) Network features measured in the PPI. (B) Structural robustness of the PPI after random removal of nodes (gray lines) and directed removal of genes ranked by mutational intolerance score (inset shows the pattern around the 20% of nodes removal, corresponding to the intolerant gene sets). (C) Deviation (z score) of every network feature. (D) Deviation (z score) of the functional features measured for every gene set (significance:  $P < 0.05 = \text{n.s.}$ ,  $0.001 < P < 0.05 = *$ ,  $0.0001 < P < 0.001 = **$ ,  $P < 0.0001 = ***$ ).

whether associations are consistent in genes affecting organismal or cellular fitness. We analyzed two features of evolutionary conservation: gene orthology and paralogy.

First, we evaluated the degree of gene conservation by calculating a gene conservation index (CI) that measures the proportion of species in which a human gene has a one-to-one ortholog (Figure 4A). We considered a total of 187 species from 7 taxonomic groups (archaea, bacteria, protozoa, fungi, plants, invertebrates and vertebrates) (Materials and Methods section). Intolerant genes are significantly

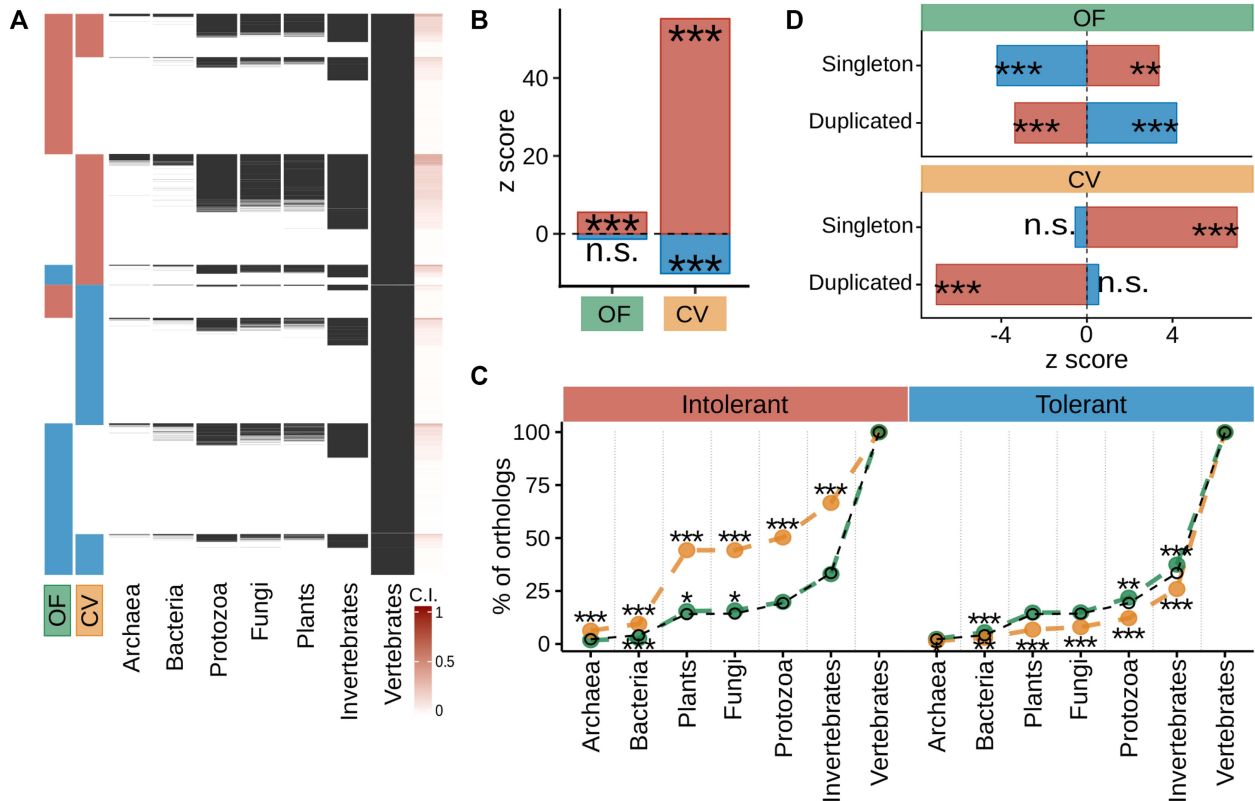
more conserved than tolerant genes in both OF and CV contexts ( $p\text{val} < 0.0001$ , two-sided z test) (Figure 4B and Supplementary Figure S2). Notably, however, the conservation of intolerant genes that affect cell viability is considerably higher than that of genes affecting organismal fitness. To further explore the difference in conservation, we calculated the proportion of genes having a one-to-one ortholog by taxonomic and tolerance group (Figure 4C). This analysis revealed a clear difference between CV and OF gene groups. In particular, CV intolerant genes are more represented in every taxonomic group except for vertebrates, while the behavior of OF intolerant genes does not deviate from random expectation, presenting only a marginal enrichment among Plants and Fungi and depletion in Bacteria (Figure 4C). This result demonstrates deep conservation of intolerant genes affecting cellular viability in humans, possibly reflecting the relevance of such genes in core cell-autonomous functions.

We next analyzed the association between tolerance groups and copy number variation, considering the number of gene duplication events represented in each gene group. The number of duplication events is consistently depleted among intolerant genes in both OF and CV. In contrast, tolerant genes show over-representation of duplication events, but only in the OF context. These results are consistent with a scenario in which paralogs might be buffering phenotypic effects of gene deletion (51) (Figure 4D). The evolutionary pattern of reduced duplication events in intolerant genes is also consistent with a reduction in gene family size distribution for intolerant relative to tolerant genes (Supplementary Figure S2). Together, these results confirm that there is a marked difference in the evolutionary history of tolerant versus intolerant genes. Within tolerant classes, we also found differences between genes that affect cell viability (CV) or organismal fitness (OF): CV intolerant genes are evolutionarily older than OF genes, and only OF but not CV tolerant genes tend to keep multiple gene copies in evolutionary history, suggesting that only organismal but not cellular fitness captures a role for paralogs on phenotypic buffering—organismal.

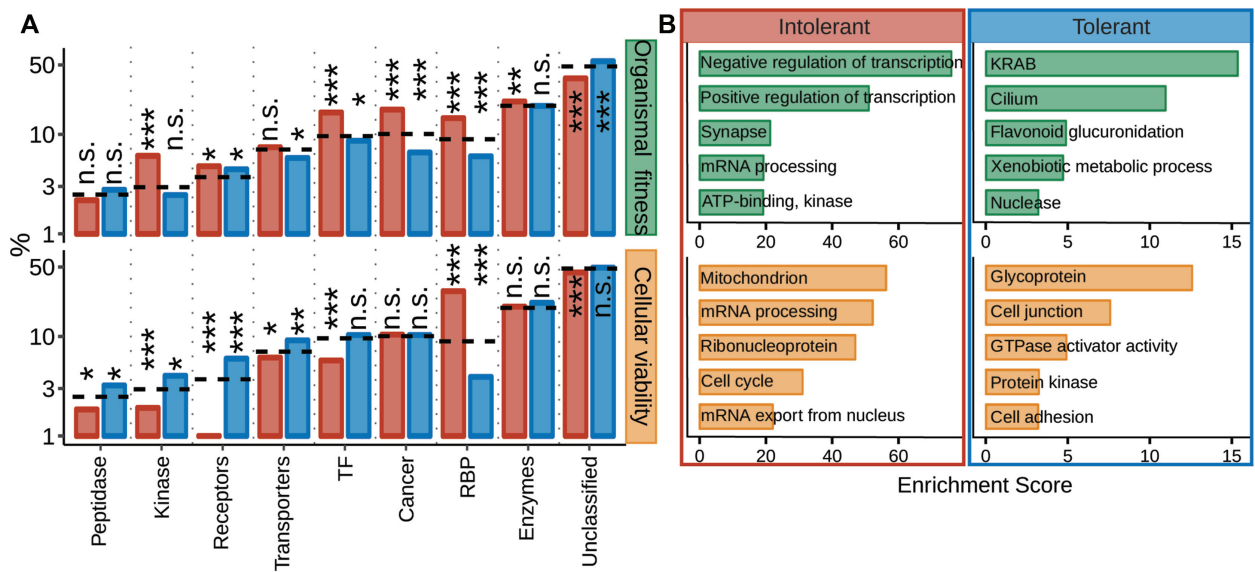
### Molecular classes predict context-dependent mutational tolerance

Our previous results revealed differences in the evolutionary history of tolerance gene classes that affect CV or OF, suggesting that the deep conservation of intolerant genes with effect in cellular viability possibly stems from their role in core unicellular functions. To further explore this association and unravel the differences between OF and CV gene sets, we analyzed the differential over-representation of tolerance gene groups within protein classes and gene ontology terms.

We again found differences in the protein class distribution of (in)tolerant genes, depending on the context of influence CV or OF (Figure 5A). Protein kinases, receptors, transcription factors (TF) and cancer associated proteins show an unexpected contrasting pattern in CV versus OF context, with enrichment of OF intolerant genes but depletion of CV intolerant genes, and a reverse pattern for tolerant genes: enrichment for CV and depletion for OF. Thus,



**Figure 4.** Genes evolutionary features. (A), Presence of gene orthologs among taxa, each row shows the presence (black) of an ortholog in the given taxon. (B) Deviation of mean gene C.I. (C) Percent of genes with an ortholog in each taxon (random expectation is shown in black). (D) Deviation in the number of singleton and duplicated genes per set. (significance:  $P < 0.05 = \text{n.s.}$ ,  $0.001 < P < 0.05 = *$ ,  $0.0001 < P < 0.001 = **$ ,  $P < 0.0001 = ***$ ).



**Figure 5.** Protein class distribution and gene ontology term enrichment. (A) Percent of genes belonging to each protein class distribution, dashed horizontal lines indicate the random expectation, deviation from expectation is shown on top of each bar (significance:  $P < 0.05 = \text{n.s.}$ ,  $0.001 < P < 0.05 = *$ ,  $0.0001 < P < 0.001 = **$ ,  $P < 0.0001 = ***$ ). (B) Top five enriched terms from functional cluster enrichment (DAVID) for each gene set.

these four protein categories, which together have a key role in developmental processes and associated signaling pathways, tend as a group to not tolerate LoF mutations in the human population, yet are not strongly required for human cell viability.

Gene ontology enrichment analysis further supports the difference between tolerance groups depending on CV or OF context, with distinct over-represented terms (Figure 5B). Consistent with the previous result, OF intolerant genes show over-representation of gene ontology terms related to development and cell communication, such as transcription regulation, kinases and synapse. CV intolerant genes, on the contrary, show over-representation of core cellular processes related to cell energetics, replication, transcription and translation. Consistent with contrasting functional properties of mutational gene tolerance classes depending on context, human genes that tolerate LoF mutations in cell culture are over-represented in processes related to cell adhesion and communication, i.e. in processes that do not tolerate mutations in the human population context (OF) (Figure 5B).

### Contrasting tissue-specificity and developmental activity of (in)tolerant genes

While global network structural and molecular functional properties provide evidence of consistent strong functional constraint on genes that do not tolerate LoF mutations in either human populations (OF context) or in human culture experiments (CV context); more detailed analyses of evolutionary history, protein classes and gene ontology terms suggest that the two contexts (OF and CV) capture distinct functional roles of intolerant genes in the organism. To further explore the hypothesis of contrasting functional constraints, we gathered and interrogated data informative of developmental involvement and tissue-specific expression.

First, we evaluated the distribution of developmental stages in which genes are first expressed. Intolerant genes are expressed earlier in development than tolerant genes in both OF and CV contexts (Figure 6A), with at least 98% of the genes already expressed in prenatal stages. On the contrary, tolerant genes are depleted in prenatal stages and preferentially expressed after birth. This similar pattern of early expression is consistent with the high involvement of both TF-mediated specification, cell-attachment and core cellular replication in embryogenesis and organogenesis. However, when considering curated gene sets involved in specific developmental processes, we found a contrasting pattern between OF and CV gene sets, consistent with previous results. In OF context, intolerant genes are over-represented in every developmental category, while tolerant genes are depleted in all categories. In sharp contrast, in CV context, tolerant genes are over-represented in developmental processes, while intolerant genes are depleted (Figure 6B). This result further supports the view that genes intolerant of LoF mutations in the human population are preferentially involved in organismal development, and that the same constraint is not captured by *in vitro* screens of gene essentiality.

In addition to developmental-stage associations, we next explored whether (in)tolerant gene classes of CV or OF context recover distinct preferential behavior in adult tis-

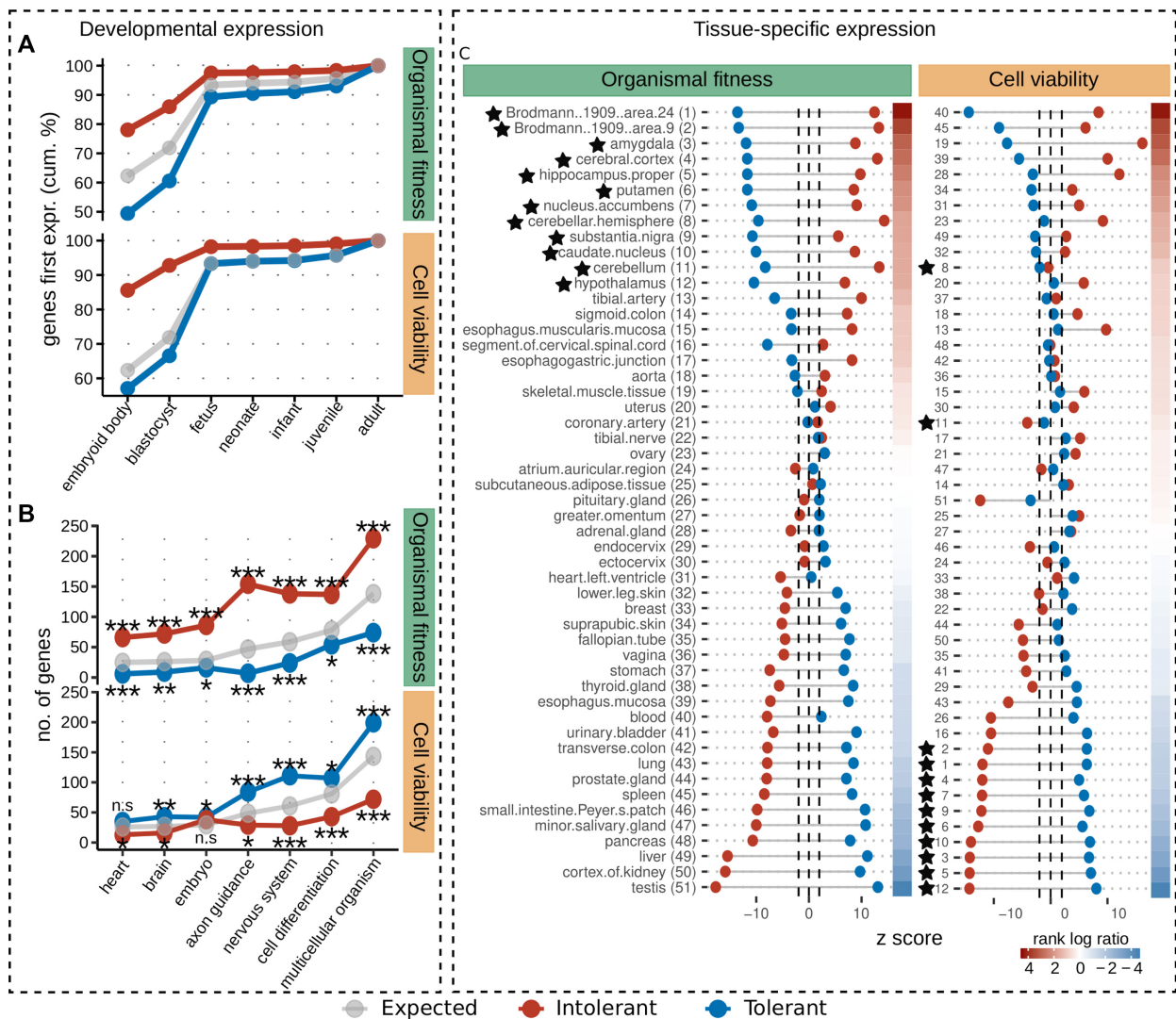
ues. We used RNA-seq data from the Genotype-Tissue Expression (GTEx) project (25) to analyze patterns of tissue-specific expression. First we performed gene expression specificity analysis to compute for each tissue and gene a quantitative measure of specific expression relative to all other tissues (Materials and Methods section). We then used the specificity values to estimate the degree to which a gene tolerance group shows unexpectedly high preferential expression in a given tissue relative to random expectation. We performed these calculations independently for each tissue and tolerance group, and for each context (CV or OF). We found both common and particular patterns of behavior among OF and CV contexts. We found consistent opposite behavior in tissue preference in tolerant versus intolerant genes: in both contexts tissues with preferential expression of intolerant genes show depleted preferential expression of tolerant genes, and vice versa (Figure 6C).

Next, to contrast the tissue-preference behavior of tolerance groups in each context, we ordered the tissues by their relative preference to preferentially express intolerant versus tolerant genes. We measured this preference by the ratio of how each tissue ranks in intolerant versus tolerant preferential expression. Using this approach, tissues that tend to preferentially express intolerant genes and not tolerant genes appear on top (Figure 6C). Notably, this analysis uncovered a contrasting behavior between OF and CV contexts: tissues from the central nervous system as a group ( $n = 12$  brain regions) show the largest relative preference of intolerant gene expression in the OF context and the least preference in the CV context. In other words, we found that the adult human brain tends to preferentially express genes that do not tolerate LoF mutations in the human population while preferentially repressing both OF tolerant genes and genes required for cell viability (CV intolerant genes) (Figure 6C). The other tissues do not show any clear pattern distinguishing CV and OF measures. We corroborated the reproducibility of these results by using independent gene expression reference data from the human protein atlas (26) (Supplementary Figure S3).

### Genes with inconsistent mutational tolerance behavior

The contrasting patterns found in OF versus CV gene groups suggests that genes with an inconsistent tolerance behavior across contexts might be driving the observed functional differences. To test this hypothesis and to identify specific genes that capture the differential functional constraints accessible through population-based versus CRISPR-bases essentiality estimations, we defined (in)consistency mutational tolerance classes and repeated all association analyses for the new groups (Figure 7A). We identified a group of 567 genes that do not tolerate LoF mutations in human populations, but that are not required for survival in human cells (organismal intolerant but cellular tolerant genes, OI-CT). Similarly, we identified a group of 351 genes that are cellular intolerant but organismal tolerant (CI-OT). Consistent with our previous results, OI-CT genes include major TF regulators of early cell lineage specification (e.g. SOX1, PAX6 and OLIG1), members of signaling pathways regulating these TFs (NOTCH1, NOTCH3, SMAD1), and genes encoding proteins relevant



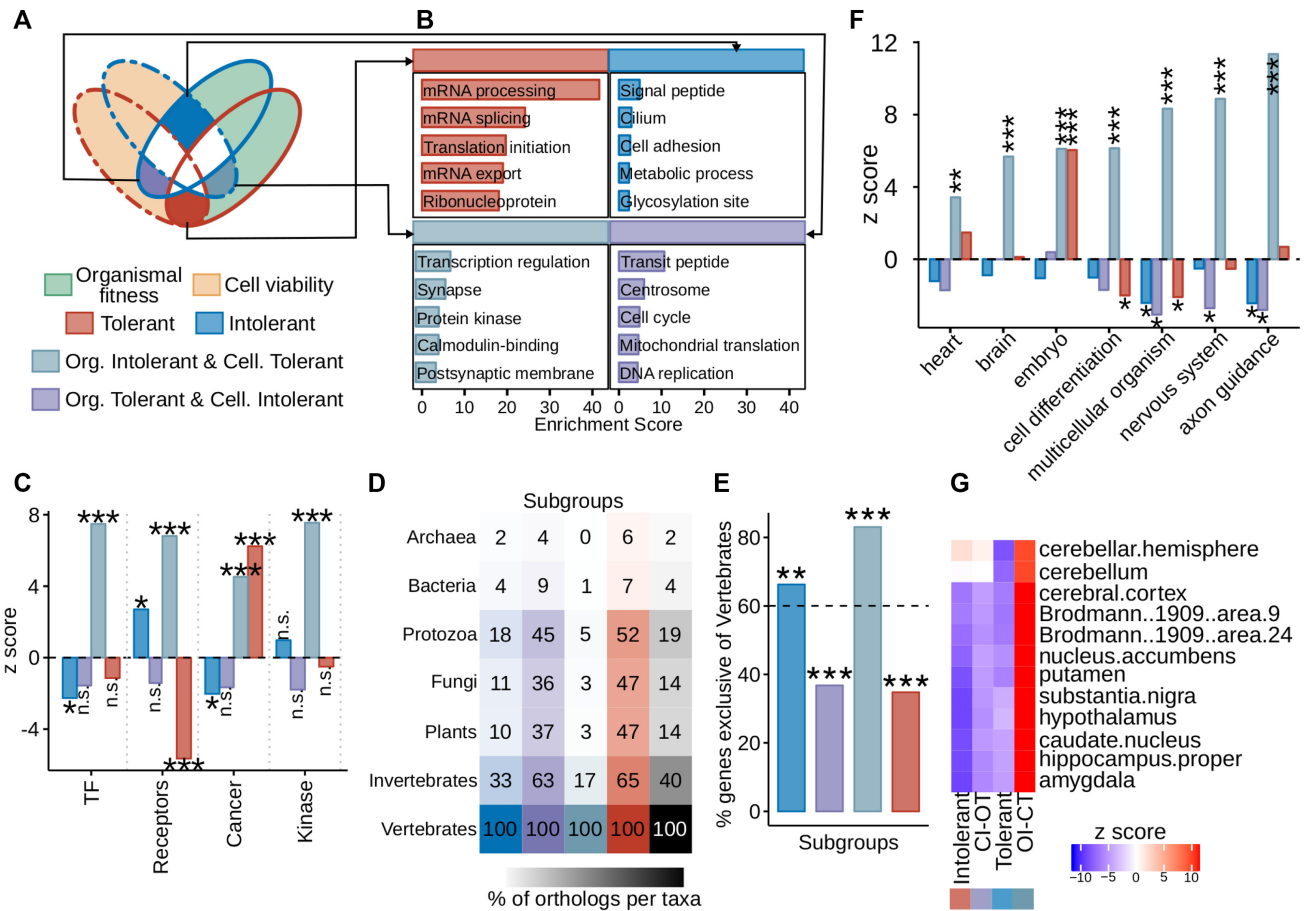


**Figure 6.** Gene set enrichment in tissue-specific expression, temporal stage expression and developmental processes. (A) Cumulative distribution of gene set percentage of genes first expressed by developmental stage. (B) Number of genes per gene set associated with a developmental process. (C) Expression enrichment deviation by tissue, tissues are ordered according to rank log ratio. Tissues highlighted with a star are part of the central nervous system (significance:  $P < 0.05 = n.s.$ ,  $0.001 < P < 0.05 = *$ ,  $0.0001 < P < 0.001 = **$ ,  $P < 0.0001 = ***$ ).

for noncell autonomous physiological integration (e.g. ion channels CACNA1C, CLCN3, GABRA1) (Supplementary Data S1).

As expected, association analyses revealed clear differences in these two inconsistent groups (OI-CT and CI-OT), in particular with respect to categories with contrasting behavior in OF versus CV gene sets. OI-CT genes are associated with gene ontology terms related to transcriptional regulation and neuronal communication, while CI-OT genes are associated with unicellular functions (Figure 7B). Similarly, protein classes with contrasting behavior in OF versus CV (i.e., TFs, receptors and kinases) are highly over-represented in OI-CT (Figure 7C). Notably, the same enrichment patterns are not observed when considering (in)tolerant genes with consistent behavior in both OF and CV contexts.

We similarly identified discrepancies in the evolutionary history of the new gene groups. OI-CT genes have less orthologs than expected within every taxonomic group except for Vertebrates, suggesting that many of these genes emerged relatively late in evolution, in pair with the emergence of vertebrates (Figure 7D). To further explore this observation, we calculated the number of genes in each tolerance gene subgroup that have one-to-one orthologs only within vertebrates, and not in other taxonomic groups. This analysis confirmed that >80% of OI-CT genes are exclusive to the vertebrate branch, in sharp contrast with both consistently intolerant genes and with CI-OT genes (Figure 7E). Lastly, OI-CT genes are also highly over-represented within every developmental process considered (Figure 7F) and are preferentially expressed in adult human brain tissues (Figure 7G). The evolutionary and functional patterns



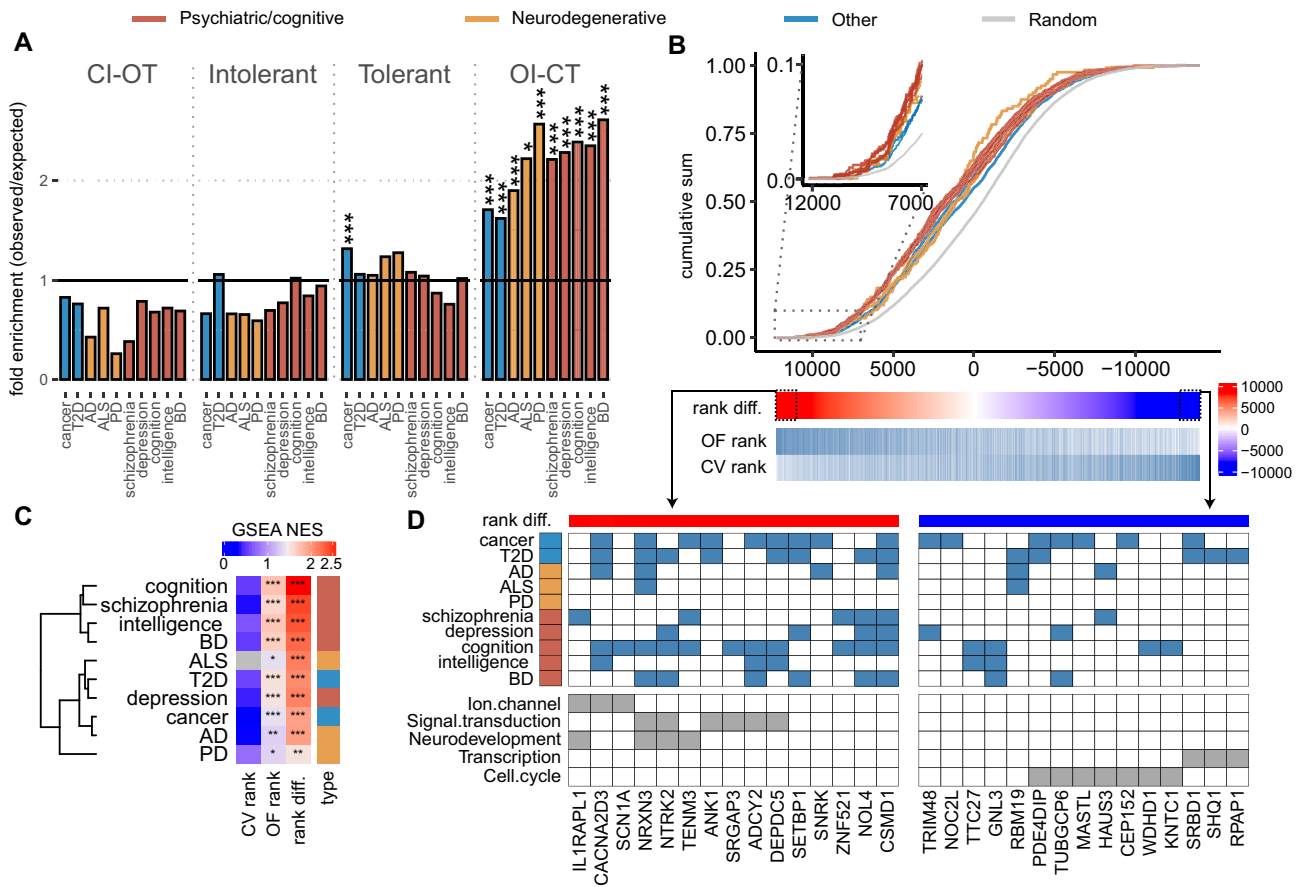
**Figure 7.** Consistent and inconsistent subgroups analyses. (A) Subgroups definitions based on the overlap between OF and CV gene sets. (B) Ontology term enrichment for each subgroup. (C) Protein class distribution enrichment. (D) Percent of orthologs in each taxon, color of the column indicates subgroup, black column shows the expected values. (E) Percent of genes that only have orthologs among vertebrates. (F) Enrichment of gene presence in developmental processes. (G) Brain tissues expression enrichment (significance:  $P < 0.05 = \text{n.s.}$ ,  $0.001 < P < 0.05 = *$ ,  $0.0001 < P < 0.001 = **$ ,  $P < 0.0001 = ***$ ).

associated with the OI-CT group suggests that the genes in this group are relevant for organismal physiology.

**Mutational intolerant genes nonessential for cell survival are associated with brain disease and cognitive traits**

Because OI-CT genes are evolutionarily novel, involved in development, preferentially expressed in the brain, and do not tolerate deleterious mutations in the human population; we hypothesized that their function might be associated with phenotypic traits characteristic of humans. To test this hypothesis, we compiled genes associated with cognitive traits (cognition and intelligence), psychiatric disorders (schizophrenia, depression and bipolar disorder BD) and neurodegenerative diseases (Alzheimer’s disease [AD], amyotrophic lateral sclerosis [ALS] and Parkinson’s disease [PD]). These and related traits and diseases have been considered either of particular relevance for human biology or human-specific (52–54). For contrast, we included cancer and type 2 diabetes (T2D), diseases not directly associated with cognitive traits. In support of our hypothesis, we found that every disease gene set is over-represented in the OI-CT genes, with the psychiatric/cognitive traits having higher

fold enrichment ( $FE > 2$  for every gene set) than the other traits (Figure 8A). In almost every other tolerant class disease associated genes are underrepresented, the only exception being consistently tolerant genes, where cancer genes are over-represented. This pattern is consistent with the relevance of OI-CT gene function in organismal physiology, and in cognitive functions and human brain neurophysiology in particular. We further explored these associations by investigating more globally the degree to which large differences in the degree to which genes tolerate detrimental mutation in the human population versus in cell culture conditions tend to recover physiologically relevant genes. To this end, we scored every gene by the rank difference between its OF and CV tolerance scores and tested whether GWAS trait associated genes tend to have unexpectedly large rank differences (Figure 8B and C). We found that, indeed, GWAS genes have large OF-CV rank differences, with stronger enrichment than that obtained when considering OF or CV tolerance scores alone, with the latter lacking any significant association. Among GWAS traits, cognitive and psychiatric traits showed the highest association with OF-CV rank differences, indicating that genes associated with these traits are particularly intolerant of detrimental



**Figure 8.** Subgroups association with psychiatric disorders and cognitive traits. (A) Over-representation of GWAS associated genes and (in)consistent tolerance groups. Fold enrichment (FE) of observed over expected overlap. Gene sets are colored according to the trait classification. (B) Cumulative sum of genes belonging to each GWAS set, genes sorted by the difference in OF and CV mutational tolerance rank. (C) Normalized enrichment score (NES) from gene set enrichment analysis (GSEA) performed test over-representation of GWAS genes with respect to CV, OF, rank difference scores. (D) Top/bottom 15 genes with extreme difference in OF and CV mutational tolerance scores and their traits and functional annotations. (significance:  $P > 0.05 = \text{n.s.}$ ,  $0.001 < P < 0.05 = *$ ,  $0.0001 < P < 0.001 = **$ ,  $P < 0.0001 = ***$ ).

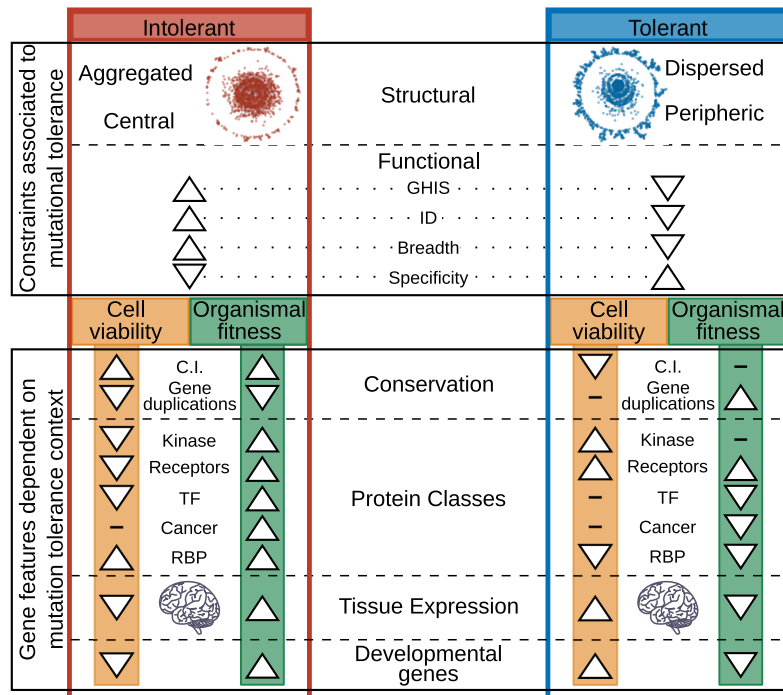
mutation in the human population and yet nonessential for cell survival (Figure 8C). To examine whether gene function might help explain such a pattern, we looked more closely at the GWAS genes having the most extreme top/bottom OF-CV rank differences (Figure 8D). Genes with high OF and low CV ranking, and thus mutationally constrained in humans but not in cells, tend to be involved in neurodevelopment and signal transduction, and to function as ion channels; processes relevant for multicellular communication and proper brain structure/function. On the other extreme, genes with high CV and low ranking, and thus mutationally constrained in cells but not in humans, are associated with cell autonomous functions like cell replication and transcription. This evidence demonstrates that genes with contrasting mutational tolerance behavior are indeed associated with cognitive traits, which are biologically relevant functions with a more recent evolutionary history.

## DISCUSSION

We examined the degree to which measures that rank human genes according to their degree of tolerance to LoF mutations capture functional constraints. We considered

tolerance estimations based on either *in vivo* exome-based population data or *in vitro* CRISPR-based perturbation experiments. To interpret evidence of differences in functional constraint in essential versus mutational tolerant genes, we integrated genome-wide data related to gene function, including structural, functional and evolutionary features. Our results indicate that intolerant genes (i) form a core network neighborhood in the human protein interactome, (ii) are enriched in molecular properties suggestive of functional constraint, (iii) are evolutionarily conserved and (iv) show preferential expression in specific tissues and developmental stages (Figure 9). The molecular and network properties that consistently discriminate intolerant from tolerant genes suggest that essentiality estimates based on mutational tolerance inference do recover functional constraints, irrespective of estimation context (OF or CV). However, we also found differences in the discriminatory properties of genes depending on whether their tolerance to mutation was estimated at the organismal or cellular level.

Consistent with previous observations (6,14), we found that structural network properties consistently discriminate tolerance/essentiality classes. Intolerant genes are central



**Figure 9.** Results summary. Enrichment patterns of the main features found associated with genes essentiality. Top panel: Structural and functional features distinctive of tolerant/intolerant genes irrespective of the context. Bottom panel: Distinctive features that show a divergent pattern depending on the context in which mutational tolerance is defined.

and localized in the human interactome, while tolerant genes are dispersed in the periphery. This relative organization predicts preferential vulnerability of the cell to intolerant gene failure, a principle that we confirmed by simulated network perturbation analysis. Both centrality and perturbation results provided results consistent with the hypothesis of a dominant role of intolerant genes in influencing cell behavior. Molecularly, intolerant genes also show properties often associated with gene functional relevance. We observe a significant tendency ( $FDR < 0.01$ ) for intolerant genes to be haploinsufficient, to have structural disorder and to be broadly expressed. In contrast, tolerant genes show under-representation of these properties. These observations further support the functional relevance of intolerant genes, as similar to broadly expressed (55), intrinsically disordered proteins are highly pleiotropic given their structural flexibility and interaction promiscuity (56), while dosage alteration of haploinsufficient genes is similarly prone to be detrimental (57).

From an evolutionary perspective, we also observe a clear distinction between tolerant and intolerant genes, consistent with previous reports (48,50). Genes intolerant to LoF mutations have an older evolutionary history, with deep one-to-one orthology across species. Notably, we found that this evolutionary pattern is accentuated in intolerant genes estimated at the cellular-level compared to those measured at the organismal-level, suggesting that CV intolerant genes are more prone to be involved in basic cell-autonomous processes shared among all taxa, with prominent presence in unicellular organisms. In contrast, OF intolerant genes

are either not enriched in unicellular groups (Archaea and Protozoa) or are underrepresented in Bacteria, suggesting that these genes emerged more recently in evolution and acquired a central role at a higher level of multicellular organization.

The idea that the context at which mutational tolerance is estimated discriminates genes operating at different levels of organization is reinforced when considering the differences we found in molecular classes among tolerance groups. The ontology terms associated with CV intolerant genes (e.g. mitochondria, RNA processing, ribonucleoprotein and cell cycle) relate to core functions required for cell survival, such as cellular metabolism and replication. On the contrary, CV tolerant genes relate to intercellular adhesion and communication (glycoprotein, cell junction and protein kinase). In sharp contrast, OF intolerant genes are enriched in functional features key to multicellularity, as evidenced by over-representation of transcriptional regulators, synapse genes, kinases and receptors. These results suggest that OF measures recover functional constraints stemming from multicellularity and organismal regulation, a property not readily captured by CV estimations. Consistent with this view, by considering curated gene annotations for developmental processes, we found that genes involved in developmental processes are enriched for OF intolerant genes and CV tolerant genes, and depleted in OF tolerant genes and CV intolerant genes.

Contrasting behaviors also manifest in tissue-specificity. We found that OF intolerant and CV tolerant genes are

preferentially expressed in the adult human brain, in contrast with the underexpression of both OF tolerant genes and genes required for cell viability (CV intolerant genes). The over-representation of OF intolerant genes in the adult brain requires a careful explanation, considering additional functions of these pleiotropic genes in the organism and close relatives, and potential reasons why they might be structurally constrained, a study beyond the scope of the current paper. From the current results, we speculate that, in addition to multicellular functional constraints, the depletion of LoF mutations estimated in human populations might capture constraints stemming from functional properties of species-specific relevance, such as higher cognition and associated traits grounded on the complexified human brain (54).

Although essentiality estimates from both cellular and organismal contexts do recover functional constraints, we found that a subgroup of 567 genes estimated as essential at the organismal level, yet nonessential at the cellular level, is responsible for the contrasting functional patterns found between OF and CV intolerant genes. These genes, which we refer to as (*OrgEssential*), are enriched in developmental processes, transcriptional regulation and neuronal communication and are preferentially expressed in the human brain. Furthermore, these genes are also associated with cognitive and psychiatric traits, underscoring the functional relevance of human-specific constraints recovered only from the organismal level of mutational intolerance. Despite being evolutionary younger than other essential genes, sharing one-to-one orthologs mainly with Vertebrates, *OrgEssential* genes seem to have developed a central role in the organism, providing an example of how during evolution novel genes can acquire essential properties by acting at levels of biological organization beyond core cell functionality.

## SUPPLEMENTARY DATA

Supplementary Data are available at NARGAB Online.

## ACKNOWLEDGEMENTS

J.C.P. is a doctoral student from Programa de Doctorado en Ciencias Biomédicas, Universidad Nacional Autónoma de México (UNAM) and received fellowship 446988 from CONACYT.

## FUNDING

Consejo Nacional de Ciencia y Tecnología (CONACYT) [46988 to J.C.P.]. Funding for open access charge: Universidad Nacional Autónoma de México (UNAM-DGAPA-PAPIIT <http://dgapa.unam.mx/index.php/impulso-a-la-investigacion/papiit>) [IN211721 to E.R.A.B.].

Conflict of interest statement. None declared.

## REFERENCES

- Mayr, E. (1964) The determinants and evolution of life. The evolution of living systems. *Proc. Natl. Acad. Sci. USA*, **51**, 934–941.
- Dobzhansky, T. and Levene, H. (1948) Genetics of natural populations; proof of operation of natural selection in wild populations of *Drosophila pseudoobscura*. *Genetics*, **33**, 537–547.
- Waddington, C.H. (1942) Canalization of development and the inheritance of acquired characters. *Nature*, **150**, 563–565.
- Gibson, G. and Dworkin, I. (2004) Uncovering cryptic genetic variation. *Nat. Rev. Genet.*, **5**, 681–690.
- Zhan, T. and Boutros, M. (2016) Towards a compendium of essential genes - From model organisms to synthetic lethality in cancer cells. *Crit. Rev. Biochem. Mol. Biol.*, **51**, 74–85.
- Bartha, I., di Iulio, J., Craig Venter, J. and Telenti, A. (2017) Human gene essentiality. *Nat. Rev. Genet.*, **19**, 51–62.
- Rancati, G., Moffat, J., Typas, A. and Pavelka, N. (2018) Emerging and evolving concepts in gene essentiality. *Nat. Rev. Genet.*, **19**, 34–49.
- Chen, P., Wang, D., Chen, H., Zhou, Z. and He, X. (2016) The nonessentiality of essential genes in yeast provides therapeutic insights into a human disease. *Genome Res.*, **26**, 1355–1362.
- Liu, G., Yong, M.Y.J., Yurieva, M., Srinivasan, K.G., Liu, J., Lim, J.S.Y., Poidinger, M., Wright, G.D., Zolezzi, F., Choi, H. *et al.* (2015) Gene essentiality is a quantitative property linked to cellular evolvability. *Cell*, **163**, 1388–1399.
- Blomen, V.A., Májek, P., Jae, L.T., Bigenzahn, J.W., Nieuwenhuis, J., Staring, J., Sacco, R., van Diemen, F.R., Olk, N., Stukalov, A. *et al.* (2015) Gene essentiality and synthetic lethality in haploid human cells. *Science*, **350**, 1092–1096.
- Wang, T., Birsoy, K., Hughes, N.W., Krupczak, K.M., Post, Y., Wei, J.J., Lander, E.S. and Sabatini, D.M. (2015) Identification and characterization of essential genes in the human genome. *Science*, **350**, 1096–1101.
- Wang, T., Wei, J.J., Sabatini, D.M. and Lander, E.S. (2014) Genetic screens in human cells using the CRISPR-Cas9 system. *Science*, **343**, 80–84.
- Wang, T., Yu, H., Hughes, N.W., Liu, B., Kendirli, A., Klein, K., Chen, W.W., Lander, E.S. and Sabatini, D.M. (2017) Gene essentiality profiling reveals gene networks and synthetic lethal interactions with oncogenic ras. *Cell*, **168**, 890–903.
- Lek, M., Karczewski, K.J., Minikel, E.V., Samocha, K.E., Banks, E., Fennell, T., O'Donnell-Luria, A.H., Ware, J.S., Hill, A.J., Cummings, B.B. *et al.* (2016) Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, **536**, 285–291.
- Kirschner, M.W., Gerhart, J.C. and Norton, J. (2005) *The Plausibility of Life: Resolving Darwin's Dilemma*. Yale University Press, New Haven, Connecticut
- Petrovski, S., Wang, Q., Heinzen, E.L., Allen, A.S. and Goldstein, D.B. (2013) Genic intolerance to functional variation and the interpretation of personal genomes. *PLoS Genet.*, **9**, e1003709.
- Rackham, O.J.L., Shihab, H.A., Johnson, M.R. and Petretto, E. (2015) EvoTo: a protein-sequence based evolutionary intolerance framework for disease-gene prioritization. *Nucleic Acids Res.*, **43**, e33.
- Samocha, K.E., Robinson, E.B., Sanders, S.J., Stevens, C., Sabo, A., McGrath, L.M., Kosmicki, J.A., Rehnström, K., Mallick, S., Kirby, A. *et al.* (2014) A framework for the interpretation of de novo mutation in human disease. *Nat. Genet.*, **46**, 944–950.
- Fadista, J., Oskolkov, N., Hansson, O. and Groop, L. (2017) LoFtool: a gene intolerance score based on loss-of-function variants in 60 706 individuals. *Bioinformatics*, **33**, 471–474.
- Bartha, I., Rausell, A., McLaren, P.J., Mohammadi, P., Tardaguila, M., Chaturvedi, N., Fellay, J. and Telenti, A. (2015) The characteristics of heterozygous protein truncating variants in the human genome. *PLoS Comput. Biol.*, **11**, e1004647.
- Cassa, C.A., Weghorn, D., Balick, D.J., Jordan, D.M., Nusinow, D., Samocha, K.E., O'Donnell-Luria, A., MacArthur, D.G., Daly, M.J., Beier, D.R. *et al.* (2017) Estimating the selective effects of heterozygous protein-truncating variants from human exome data. *Nat. Genet.*, **49**, 806–810.
- Hart, T., Chandrashekar, M., Aregger, M., Steinhart, Z., Brown, K.R., MacLeod, G., Mis, M., Zimmermann, M., Fradet-Turcotte, A., Sun, S. *et al.* (2015) High-Resolution CRISPR screens reveal fitness genes and genotype-specific cancer liabilities. *Cell*, **163**, 1515–1526.
- Dosztanyi, Z., Csizmek, V., Tompa, P. and Simon, I. (2005) IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics*, **21**, 3433–3434.
- Steinberg, J., Honti, F., Meader, S. and Webber, C. (2015) Haploinsufficiency predictions without study bias. *Nucleic Acids Res.*, **43**, e101–e101.

25. Carithers,L.J., Ardlie,K., Barcus,M., Branton,P.A., Britton,A., Buia,S.A., Compton,C.C., DeLuca,D.S., Peter-Demchok,J., Gelfand,E.T. *et al.* (2015) A novel approach to High-Quality postmortem tissue Procurement: The GTEx project. *Biopreserv. Biobank.*, **13**, 311–319.
26. Uhlen,M., Oksvold,P., Fagerberg,L., Lundberg,E., Jonasson,K., Forsberg,M., Zwahlen,M., Kampf,C., Wester,K., Hober,S. *et al.* (2010) Towards a knowledge-based human protein atlas. *Nat. Biotechnol.*, **28**, 1248–1250.
27. Kryuchkova-Mostacci,N. and Robinson-Rechavi,M. (2017) A benchmark of gene expression tissue-specificity metrics. *Brief. Bioinform.*, **18**, 205–214.
28. Lambert,S.A., Jolma,A., Campitelli,L.F., Das,P.K., Yin,Y., Albu,M., Chen,X., Taipale,J., Hughes,T.R. and Weirauch,M.T. (2018) The human transcription factors. *Cell*, **175**, 598–599.
29. Gerstberger,S., Hafner,M. and Tuschl,T. (2014) A census of human RNA-binding proteins. *Nat. Rev. Genet.*, **15**, 829–845.
30. Southan,C., Sharman,J.L., Benson,H.E., Faccenda,E., Pawson,A.J., Alexander,S.P.H., Buneman,O.P., Davenport,A.P., McGrath,J.C., Peters,J.A. *et al.* (2016) The IUPHAR/BPS Guide to PHARMACOLOGY in 2016: towards curated quantitative interactions between 1300 protein targets and 6000 ligands. *Nucleic Acids Res.*, **44**, D1054–D1068.
31. Herrero,J., Muffato,M., Beal,K., Fitzgerald,S., Gordon,L., Pignatelli,M., Vilella,A.J., Searle,S.M.J., Amode,R., Brent,S. *et al.* (2016) Ensembl comparative genomics resources. *Database*, **2016**.
32. Chen,W.-H., Minguez,P., Lercher,M.J. and Bork,P. (2012) OGEE: an online gene essentiality database. *Nucleic Acids Res.*, **40**, D901–D906.
33. Chen,W.-H., Lu,G., Chen,X., Zhao,X.-M. and Bork,P. (2017) OGEE v2: an update of the online gene essentiality database with special focus on differentially essential genes in human cancer cell lines. *Nucleic Acids Res.*, **45**, D940–D944.
34. Li,T., Wernersson,R., Hansen,R.B., Horn,H., Mercer,J., Slodkowitz,G., Workman,C.T., Rigina,O., Rapacki,K., Staerfeldt,H.H. *et al.* (2017) A scored human protein–protein interaction network to catalyze genomic interpretation. *Nat. Meth.*, **14**, 61–64.
35. Cornish,A.J. and Markowitz,F. (2014) SANTA: quantifying the functional content of molecular networks. *PLoS Comput. Biol.*, **10**, e1003808.
36. Gabor,C. and Nepusz,T. (2006) The igraph software package for complex network research. *InterJ. Complex Syst.*, **1695**, 1–9.
37. Vilella,A.J., Severin,J., Ureta-Vidal,A., Heng,L., Durbin,R. and Birney,E. (2009) EnsemblCompara Genetrees: Complete, duplication-aware phylogenetic trees in vertebrates. *Genome Res.*, **19**, 327–335.
38. Ritchie,M.E., Phipson,B., Wu,D., Hu,Y., Law,C.W., Shi,W. and Smyth,G.K. (2015) limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.*, **43**, e47.
39. Huang,D.W., Sherman,B.T. and Lempicki,R.A. (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.*, **4**, 44–57.
40. Huang,D.W., Sherman,B.T., Zheng,X., Yang,J., Imamichi,T., Stephens,R. and Lempicki,R.A. (2009) Extracting biological meaning from large gene lists with DAVID. *Curr. Protoc. Bioinform.*, **27**, 1–13.
41. Buniello,A., MacArthur,J.A.L., Cerezo,M., Harris,L.W., Hayhurst,J., Malangone,C., McMahon,A., Morales,J., Mountjoy,E., Sollis,E. *et al.* (2019) The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.*, **47**, D1005–D1012.
42. Wang,M., Zhao,Y. and Zhang,B. (2015) Efficient test and visualization of Multi-Set intersections. *Sci. Rep.*, **5**, 16923.
43. Korotkevich,G., Sukhov,V., Budin,N., Shpak,B., Artyomov,M.N. and Sergushichev,A. (2021) Fast gene set enrichment analysis. bioRxiv doi: <https://doi.org/10.1101/060012>, 01 February 2021, preprint: not peer reviewed.
44. UniProt Consortium (2021) UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res.*, **49**, D480–D489.
45. Zotenko,E., Mestre,J., O’Leary,D.P. and Przytycka,T.M. (2008) Why do hubs in the yeast protein interaction network tend to be Essential: Reexamining the connection between the network topology and essentiality. *PLoS Comput. Biol.*, **4**, e1000140.
46. Batada,N.N., Hurst,L.D. and Tyers,M. (2006) Evolutionary and physiological importance of hub proteins. *PLoS Comput. Biol.*, **2**, e88.
47. Yu,H., Greenbaum,D., Lu,H.X., Zhu,X. and Gerstein,M. (2004) Genomic analysis of essentiality within protein networks. *Trends Genet.*, **20**, 227–231.
48. Fraser,H.B. (2002) Evolutionary rate in the protein interaction network. *Science*, **296**, 750–752.
49. Fraser,H.B. (2005) Modularity and evolutionary constraint on proteins. *Nat. Genet.*, **37**, 351–352.
50. Hahn,M.W. and Kern,A.D. (2005) Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks. *Mol. Biol. Evol.*, **22**, 803–806.
51. Gu,Z., Steinmetz,L.M., Gu,X., Scharfe,C., Davis,R.W. and Li,W.-H. (2003) Role of duplicate genes in genetic robustness against null mutations. *Nature*, **421**, 63–66.
52. Burns,J.K. (2004) An evolutionary theory of schizophrenia: cortical connectivity, metarepresentation, and the social brain. *Behav. Brain Sci.*, **27**, 831–855.
53. Varki,A., Geschwind,D.H. and Eichler,E.E. (2008) Human uniqueness: genome interactions with environment, behaviour and culture. *Nat. Rev. Genet.*, **9**, 749–763.
54. Geschwind,D.H. and Rakic,P. (2013) Cortical evolution: judge the brain by its cover. *Neuron*, **80**, 633–647.
55. Watanabe,K., Stringer,S., Frei,O., Umičević Mirkov,M., de Leeuw,C., Polderman,T.J.C., van der Sluis,S., Andreassen,O.A., Neale,B.M. and Posthuma,D. (2019) A global overview of pleiotropy and genetic architecture in complex traits. *Nat. Genet.*, **51**, 1339–1348.
56. Wright,P.E. and Dyson,H.J. (2015) Intrinsically disordered proteins in cellular signalling and regulation. *Nat. Rev. Mol. Cell Biol.*, **16**, 18–29.
57. Huang,N., Lee,I., Marcotte,E.M. and Hurler,M.E. (2010) Characterising and predicting haploinsufficiency in the human genome. *PLoS Genet.*, **6**, e1001154.



# E Understanding epigenomics in the context of Waddington's epigenetic landscape

Artículo sometido para publicación en Bioessays.



# Understanding epigenomics in the context of Waddington's epigenetic landscape

José Luis Caldú-Primo, José Dávila-Velderrain, Juan Carlos Martínez-García,  
and Elena R. Álvarez-Buylla

## Abstract

Systemic understanding of epigenetic modifications on the chromatin could be attained by analyzing the effects they have on the topology and dynamics of the regulatory network underlying the epigenetic landscape. The understanding of cell differentiation and morphogenesis has been successfully approached through the use of the systems-based epigenetic landscape theoretical framework. The epigenetic landscape, initially proposed by Waddington, has become a powerful tool to quantitatively address constraints underlying development using gene regulatory network models. Nonetheless, given the recent understanding of gene control by epigenomic modifications the notion of epigenetics has been mainly related to non-genetic heritable modifications of the genome. Therefore, this approach has given proximal epigenomic modifications a central role in understanding development, leaving the original dynamic view of the epigenetic landscape aside. In this essay, we aim at establishing a conceptual link between both conceptualizations of epigenetics, aiming to reach a better understanding of cell differentiation and development.

## Introduction

The popularity of epigenetics among the scientific community and the lay public has strongly increased during the last decades.<sup>[1–3]</sup> From leading scientific journals to airport newsstands, the term epigenetics has found its way into popular culture in a remarkable way. We believe public interest in epigenetics can be attributed to two main reasons: i) the agreement on the fact that genetic information is not enough to understand how an organism is formed and functions and ii) the possibility brought by high-throughput next generation sequencing technologies to profile epigenomic marks (EM) on the chromatin.<sup>[4,5]</sup> As a result, the term “epigenetics” is used as a new universal explanation for virtually any biological question that cannot be explained by classical genetics. In this essay we make a distinction between two different understandings of epigenetics, namely one referring to molecular modifications of the genome and the other referring to the complex developmental processes relating genotype and phenotype. We further discuss whether the two views can be brought together under a common theoretical framework encompassing Waddington’s original conception of developmental epigenetics and modern findings of molecular EMs and gene regulation.

As a result of the historical development of epigenetics, that we briefly review in Box 1, today two interpretations of the concept coexist: i) a general one referring to the entire series of complex dynamic interactions mapping genotype and phenotype (we will refer to this as dynamical epigenetics), and ii) the study of molecular modifications on chromatin that might have a phenotypic effect independent of changes in the genetic sequence (we will refer to this as molecular epigenetics).<sup>[3,6,7]</sup> Both perspectives propose non–genetic mechanisms as the source of phenotypic state memory underlying cellular differentiation and development. Molecular epigenetics looks for memory in the form of material entities (DNA methylation, histone covalent modifications, chromatin structure, etc.). On the other hand, dynamical epigenetics is based on the realization that phenotypic memory, as a process, emerges naturally from the mutual constraints imposed by the complex regulatory relationships among genetic, cellular, organismic and environmental elements.<sup>[26, 48]</sup>

Nowadays, it is possible to characterize molecular epigenetic marks in a context dependent manner, allowing a much more precise knowledge of the proximate molecular mechanisms controlling gene expression.<sup>[5,8,9]</sup> The expansion of knowledge on mechanisms of chromatin based gene regulation and technological advances in molecular profiling techniques have given EMs a primary explanatory role of biological phenomena related to development and cell differentiation.<sup>[17, 33]</sup> It is still paradoxical that the success of Waddington’s ideas in inspiring the incursion of molecular biologists into the field of developmental biology results in the non-

1  
2  
3 use of his conceptual treatment. Still, the complexity involved in Waddington's understanding  
4 of development cannot be reduced to a mechanistic explanation that takes the form of a  
5 collection of molecular marks found on the genome. A way to move forward in the field is  
6 finding a way to include and interpret EMs from the theoretical foundation of the epigenetic  
7 landscape, updating in this way Waddington's proposal to modern knowledge on gene  
8 regulation. To discuss these ideas, in the next sections we will briefly review some major  
9 problems molecular epigenetics has when explaining development and the general basis of  
10 dynamical epigenetics.  
11  
12  
13  
14  
15

## 16 17 **Molecular epigenomic marks are not sufficient to** 18 **understand cell differentiation** 19 20 21

22  
23 The importance of EMs in cell differentiation and the maintenance of cell fate has become  
24 evident from. This has been evidenced by demonstrating that changes in cellular identity are  
25 associated with changes in the epigenomic profile and alterations in the chromatin modifying  
26 enzymes cause switches in cell fate.<sup>[4,10,11]</sup> This has led to the idea that epigenomic  
27 modifications are instructions 'written' over the DNA sequence to provide an additional level  
28 of gene regulation that in many cases is heritable.<sup>[12,13]</sup>  
29  
30  
31  
32

33  
34 With the increasing breadth of EMs, it is usual to find studies implicitly assuming that the  
35 epigenomic profile is in fact a sufficient explanation for the observed patterns of gene  
36 activation/silencing happening during cell differentiation.<sup>[14–16]</sup> These kind of explanations take  
37 for granted that the proximate molecular effectors of gene regulation (i.e. EMs) are sufficient  
38 to understand why a cell is in a given state, as illustrated in Fig. 1. Nevertheless, despite the  
39 great advances in molecular profiling, the epigenomic description of a cellular state is still  
40 insufficient to understand why the cell acquires and stays at a given phenotypic state. As  
41 stated by the Roadmap Epigenomics Consortium, "Despite these technological advances, we  
42 still lack a systematic understanding of how the epigenomic landscape contributes to cellular  
43 circuitry, lineage specification, and the onset and progression of human disease."<sup>[9]</sup> The  
44 problem being that, even though having a very detailed genomic/epigenomic profile of a given  
45 cell informs about which genes are being expressed, it does not explain how a given profile is  
46 formed and what makes it remain instead of being removed. In other words, the epigenomic  
47 profile contains information about the regulatory state of a given cell, but this information is a  
48 result of the developmental process itself.<sup>[17]</sup>  
49  
50  
51  
52  
53  
54  
55  
56  
57

58 The consideration of EMs as the explanation for cellular differentiation entails two big problems:  
59 i) most chromatin modifying enzymes (CME) responsible of placing/removing EMs lack  
60

1  
2  
3 sequence specificity and ii) EMs are highly dynamic and reversible.<sup>[3,18–20]</sup>  
4  
5

6 The absence of DNA-binding domains in CMEs responsible for DNA methylation and histone  
7 modifications makes it necessary for them to cooperate with transcription factors (TFs) to  
8 guide their activity on the genome.<sup>[12,21]</sup> The involvement of TFs seems to be a way to  
9 overcome problem of sequence specificity, but this brings up yet another problem for  
10 understanding epigenetic control of gene function: are TFs drivers and CME follow their  
11 activity placing EM on the genome, or on the contrary EMs drive TFs activity determining the  
12 places where they can bind the genome?<sup>[20,22]</sup> There is a complex relationship between TFs  
13 and EMs, with cases in which DNA methylation inhibits TF binding and other situations in  
14 which TF binding causes local loss of DNA methylation.<sup>[23,24]</sup> An example of the latter are the  
15 pioneer TFs, which can bind condensed chromatin eliciting changes in the local chromatin  
16 structure.<sup>[25]</sup> A similarly complex relationship in which there is no clear cause and effect exists  
17 between DNA methylation, histone modifications, Polycomb mediated silencing and their  
18 control over gene expression.<sup>[21,26]</sup>  
19  
20  
21  
22  
23  
24  
25  
26

27 The second problem referred to above, the reversibility of EMs, is also related to the way their  
28 presence on the chromatin is determined. As mentioned above, the spatiotemporal distribution  
29 and maintenance of EMs depends on the coordinated action of CMEs. CMEs are broadly  
30 classified in writers (e.g. histone methyltransferases and acetyltransferases), erasers (e.g.  
31 histone demethylases and deacetylases), and readers (e.g. bromodomains, zinc fingers,  
32 chromodomains). Different kinds of CMEs interact in complex regulatory circuits integrating  
33 different cellular pathways that eventually determine the placement or removal of EMs on the  
34 chromatin.<sup>[27,28]</sup> Thus, the presence of EMs on the chromatin depends on the activity of CMEs  
35 and TFs, so in order to explain their presence it is necessary to have an underlying regulatory  
36 system controlling the expression of TFs and CMEs.  
37  
38  
39  
40  
41  
42  
43

44 Despite the important role EMs have on gene regulation, the information guiding development  
45 emerges in fact from somewhere else. From a system's biology view, this information arises  
46 from the complex relationships among TFs, chromatin modifying enzymes, metabolic factors  
47 in the cytoplasm, and the whole transcriptional machinery regulating cellular behavior.<sup>[3,26,29]</sup>  
48 This view recovers the dynamical systems conception of epigenetics originally framed by  
49 Waddington, with EMs being yet another factor in the complex system underlying development.  
50 In order to integrate EMs to the conceptual framework proposed by Waddington, it is  
51 necessary to approach the way their regulatory mechanisms can be integrated into the  
52 epigenetic landscape.  
53  
54  
55  
56  
57  
58  
59  
60

## Epigenetics from a dynamical systems perspective: the epigenetic landscape

The epigenetic landscape is a diagrammatic representation of development as a multidimensional dynamical system defined by the interactions among genes and the environment. Waddington depicted his idea of a “concatenation of processes linked together in a network”<sup>[30]</sup> guiding development causing the system to reach stable end states sharply distinct from one another in his popular diagram of the ‘epigenetic landscape’.<sup>[30]</sup> Nowadays in the field of systems biology, the epigenetic landscape is considered not only an illustrative metaphor, but an experimentally and mathematically grounded theory of development.<sup>[3,31]</sup> At the center of this theory lies the gene regulatory network (GRN), a dynamical system that causally determines the activity of genes based on the regulatory interactions between gene products (especially TFs) and gene promoters controlling their transcription. In this sense, GRNs determine what genes are going to be expressed in a cell depending on the genes present at a given time (state of the system) and the regulatory interactions among them. From this perspective, the different cell types of an organism correspond to the stable states (attractors) determined in the underlying GRN and development is the motion of the cells through the attractor’s landscape. The dynamical systems modeling of the epigenetic landscape of attractors is briefly described in Box 2.

The study of organisms as dynamical systems has had great advances since Waddington introduced his creative ideas. Stuart Kauffman formalized the idea of a global GRN determining cell differentiation as a multistable attractor landscape,<sup>[32]</sup> cell types have been associated to these attractor states with gene expression levels as system state descriptions,<sup>[33]</sup> and empirical gene regulatory networks have been successfully inferred, simulating their dynamical behavior, and explaining the observed phenotypes in several developmental processes.<sup>[34,35]</sup> Still, more recent knowledge on transcriptional regulation in multicellular organisms has raised important objections on the suitability of using GRN models as something more than ‘toy models’ to understand development.<sup>[36]</sup>

One of the problems brought up by Stuart Newman is the assumption that there is a global network with a fixed topology across conditions.<sup>[36]</sup> GRN models are built from the idea of ‘gene-switches’, based on analogies with bacterial gene regulation, in which TFs have a clear and constant regulatory function across conditions. This conjecture is refuted by the fact that most TFs have intrinsically disordered regions making their structural conformation and therefore their possible interactions context dependent.<sup>[37,38]</sup> Furthermore, evidence of chromatin-based control of gene expression through EMS placed on the chromatin by CMEs

1  
2  
3 changes the available binding sites for TFs also in a context dependent manner.<sup>[20,39]</sup> TFs'  
4 intrinsic disorder and chromatin's context dependent landscape blur the existence of a global  
5 regulatory architecture underlying GRN models.  
6  
7

8  
9 Another question raised by Newman concerns the possibility for the appearance of new cell  
10 types in evolution from a given epigenetic landscape determined by the GRN. Phylogeny  
11 shows that throughout metazoan evolution there has been an increase in the number of cell  
12 types with a continuity in the ancestral cell types identity.<sup>[40]</sup> From an epigenetic landscape  
13 perspective, this process would require the GRN of the derived species having attractors  
14 corresponding to the new cell types, while keeping the ancestral attractors maintaining their  
15 compositional identity and the hierarchical relationships among them. The maintenance of a  
16 system's attractors structure when adding new variables and generating new attractors is not  
17 a general property of dynamical systems, as such cell type evolution is not easily explained  
18 from the classical epigenetic landscape framework.<sup>[36]</sup>  
19  
20  
21  
22  
23  
24  
25

26 These issues call for a reevaluation of the way the epigenetic landscape has been formally  
27 conceptualized as emerging from a hardwired GRN encoded in the genome. Chromatin based  
28 mechanisms of gene regulation and TFs structure variability make the underlying GRN highly  
29 plastic, consequently increasing the epigenetic landscape context adaptability. Chromatin  
30 based regulation adds new layers of transcriptional regulation that could allow for the  
31 appearance of new cell type attractors while maintaining the ancestral attractors structure.  
32 This idea is underscored by the fact that the generation of a higher number of cell types during  
33 metazoan evolution is associated with the appearance of new epigenomic mechanisms and  
34 distal regulatory elements.<sup>[41,42]</sup> Incorporating EMs to GRN modeling of the epigenetic  
35 landscape is a way to bring back together the two views of epigenetics under a common  
36 framework that would concurrently give a systematic understanding of EMs and update the  
37 dynamical systems modeling of development.  
38  
39  
40  
41  
42  
43  
44  
45

## 46 **Bridging the gap**

47  
48  
49 In the preceding sections we have presented two views of epigenetics that aim to explain  
50 organismic development and cell differentiation, highlighting some problems both approaches  
51 have. Now we will elaborate on ideas to integrate knowledge on chromatin based gene  
52 regulation to the classical dynamical systems modeling of the epigenetic landscape as a way  
53 to overcome some of the problems mentioned above. In particular, we hypothesize how EMs,  
54 functioning as genetic activators or silencers, can be incorporated to the theoretical conception  
55 of the epigenetic landscape, conceived as an abstraction of a multidimensional dynamical  
56  
57  
58  
59  
60

1  
2  
3 system.  
4  
5

6 First, we want to explore how context dependent epigenomic profiles can be interpreted from  
7 the epigenetic landscape perspective. Epigenomic profiling allows the inference of genome-  
8 wide coverage of EMs in a condition specific manner. Different high throughput techniques  
9 make it possible to evaluate the distribution of DNA methylation, histone modifications,  
10 chromatin accessibility, or genome architecture in different cell types or environmental  
11 conditions.<sup>[5]</sup> With this information and knowledge of the functional effects EMs have on the  
12 surrounding genes, it is possible to translate this data into chromatin states and deduce the  
13 activity states of genes.<sup>[5,9]</sup> Since the presence of EMs depends on the underlying CMEs and  
14 TFs, their distribution along the genome reflects the particular system's state. In this sense,  
15 the epigenomic profile can be interpreted as a representation of the system's state,  
16 complementary to its usual description as the genes transcriptional pattern.<sup>[33]</sup> Still, as  
17 mentioned above, given the complex interplay between EMs, TFs, and CMEs, it is difficult to  
18 determine if the transcriptional environment of a cell sets up a given epigenomic distribution,  
19 or if on the other hand a preexistent epigenomic environment determines the transcriptional  
20 activity of the cell.<sup>[12,20]</sup> This paradoxical problem might be solved through more studies on the  
21 interplay between TFs and epigenomic modifiers. Nevertheless, we strongly believe that an  
22 ultimate clear cause and consequence relationship will remain elusive. We support the idea  
23 that transcriptional and epigenomic elements of the cell have a dialectical relationship, in which  
24 their role as causes or consequences originates only through their interaction.<sup>[17,43]</sup> This means  
25 that the role a particular element plays cannot be determined *a priori*, since it is dependent on  
26 the state of the system at that particular moment. Our claim is that there is not a simple causal  
27 map between the cellular epigenomic environment and the epigenetic landscape, instead  
28 these are mutually determined through complex relationships.  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42

43 We must point out that epigenomic profiling also allows the inference of condition-specific TF  
44 regulatory interactions, this can be done by mapping chromatin accessible sites through  
45 DNase-seq, ATAC-seq, or other omics assays, and matching TFs to their available binding  
46 sites in gene promoters. Through this technique it is possible to build state-specific TF  
47 regulatory networks.<sup>[44–46]</sup> From the classical epigenetic landscape perspective, it is  
48 troublesome to speak about condition-specific regulatory networks because it is understood  
49 that the different cellular states emerge from an underlying global regulatory network. As  
50 mentioned above, the validity of conceiving a global regulatory network has been  
51 compromised by the context dependent nature of TFs structure and chromatin  
52 architecture.<sup>[36,37]</sup> Still, it is interesting to think how condition-specific networks can be  
53 understood from a dynamical systems perspective?  
54  
55  
56  
57  
58  
59  
60

1  
2  
3 If we consider the epigenomic state of a cell as a feature of the cellular state (system's attractor)  
4 as stated above, the particular condition-specific network topology inferred from these  
5 epigenomic data would also be a description of the system's state. Taking into consideration  
6 that condition-specific networks represent regulatory interactions, they can be considered a  
7 dynamical system with an associated dynamic of their own. Here it is important to think about  
8 the difference in stability and timescales between epigenomic and transcriptomic regulation.  
9 Despite EMs being reversible, in general they have a stable expression in a given cell type.<sup>[9]</sup>  
10 In fact, genome-wide reprogramming of EMs on differentiated cell types happens rarely during  
11 life, namely during fertilization and in the progression of cancer.<sup>[47,48]</sup> On the other hand,  
12 condition-specific networks represent transcriptional regulation dynamics happening at shorter  
13 timescales. What this idea points at is a process in which during development the system  
14 reaches stable states with an associated epigenomic profile, which in turn define a more  
15 constrained regulatory network by reducing the regions of the genome accessible to regulatory  
16 interactions. In this way, the developmental dynamics through which a cell acquires a given  
17 epigenomic profile, influence its underlying regulatory architecture. Accordingly, chromatin  
18 modifications are a way in which transcriptional programs are incorporated to the regulatory  
19 network, consequently becoming independent of the conditions that first brought them  
20 about.<sup>[36]</sup> Deep understanding of the interplay between transcriptional, epigenomic, and  
21 environmental factors influencing cell dynamics is an exciting challenge that will surely shed  
22 light on our understanding of development and cell differentiation.

23  
24  
25 Having in mind the ideas presented above, let's consider now how they can be applied to  
26 mammalian development. Starting from zygote fertilization, the developmental process starts  
27 with the removal of the gametic EMs from the newly formed zygote.<sup>[49,50]</sup> The process of zygotic  
28 epigenomic reprogramming and pre-implantation embryo development, a very stimulating  
29 subject of study nowadays, reaches the blastocyst stage, where embryonic stem cells are  
30 found.<sup>[51–53]</sup> Embryonic stem cells, defined by their capacity to self-renew and differentiate into  
31 any adult cell type, have a characteristic DNA organization, different from the one that  
32 characterizes adult cell types.<sup>[54,55]</sup> Their chromatin appears to be more 'open' with dispersed  
33 heterochromatin, enrichment of histone modifications associated with transcriptional activity,  
34 and reduced DNA methylation.<sup>[56,57]</sup> During embryonic development, differentiating cells  
35 acquire epigenomic modifications that 'silence' certain parts of the genome and restrict their  
36 transcriptional capacities, in fact defining them as differentiated cell-types.<sup>[55,57]</sup> Considering  
37 the ideas we introduced above and these evidences, it can be hypothesized that the step-wise  
38 acquisition of EMs during differentiation has the effect of stabilizing the systems dynamics in  
39 the adult cell types, progressively limiting the system's dynamics to the corresponding lineage  
40 at-tractors. This would add up to the directionality and the almost irreversibility of the



1  
2  
3 differentiation process, because before switching lineages or reprogramming cell fates it would  
4 be necessary to rewire the epigenomic profile.<sup>[12,22]</sup> Seen from the epigenetic landscape, the  
5 acquisition of EMs is a central characteristic of cell state canalization.  
6  
7  
8

## 9 **Conclusions**

10  
11  
12 Molecular epigenomics has expanded our knowledge on gene regulation, revealing the  
13 multiplicity and complexity of mechanisms involved in gene regulation. Nevertheless, knowing  
14 the distribution of EMs on the genome does not explain why a given cell type has that particular  
15 distribution or why different epigenomic profiles exist in the first place. As has been done so  
16 far, epigenomics has lacked the necessary conceptual tools to answer these kinds of  
17 questions. Deep understanding of the basic generic principles that causally explain biological  
18 development will hardly come solely from the study of the patterns present in databases with  
19 epigenomic information. We firmly believe that epigenomics urgently requires a systems  
20 based mechanistic perspective. The analysis of organisms as multidimensional nonlinear  
21 dynamical systems, as proposed initially by Conrad H. Waddington, is still the best theory to  
22 tackle the deep understanding of development, cellular differentiation and the reasons that  
23 causally explain the existence of different epigenomic profiles. The effects that EMs have on  
24 the dynamic network can be understood as changes in the network topology and,  
25 consequently, as modifications of the dynamics of the regulatory network, as graphically  
26 depicted in Fig. 3. In this way, the epigenomic modifications would act as a regulatory loop  
27 linking the system's transcriptional state with its regulatory dynamics. This idea could bring  
28 together and enrich the two fields of epigenetics that coexist separately these days. On one  
29 hand, it gives a systemic comprehension of epigenomic profiles by recognizing they are  
30 determined by the underlying regulatory system and as such a characteristic of the steady  
31 states of the system. On the other hand, acknowledging that EMs exert an influence over the  
32 system's network topology and dynamics gives the systems theory of development another  
33 layer of control in which the system influences its own dynamics and stability.  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48

## 49 **Acknowledgments**

50  
51 The authors declare no conflict of interest. We thank Dr. Julián Valdés for critical reading of  
52 the manuscript. J.L.C.P. is a doctoral student from Programa de Doctorado en Ciencias  
53 biomédicas, Universidad Nacional Autónoma de México (UNAM) and received fellowship  
54 446988 from CONACYT. E.R.A.B. received funding from UNAM-DGAPA-PAPIIT INN211721.  
55 J.C.M.G. received funding from CONACYT FORDECYT-PRONACES/194186/2020.  
56  
57  
58  
59  
60

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

For Peer Review

## Box 1: The origin and evolution of the term epigenetics

In the 1940s, the British embryologist Conrad H. Waddington conceived a general theory of development as a way to unify the disciplines of embryology, genetics, and evolution. During this endeavor, he coined the term 'epigenetics', from the addition of the words epigenesis and genetics, as "a suitable name for the branch of biology which studies the causal interactions between genes and their products which bring the phenotype into being".<sup>[58]</sup> As reflected by the definition, Waddington's idea of epigenetics was a broad concept encompassing "the whole complex of developmental processes" connecting genotype and phenotype.<sup>[30]</sup> At the time, molecular biology was a new branch of biology and the DNA structure and biochemistry of transcriptional regulation were still unknown. The rate of diffusion of intellectual constructions in science was much slower than what it is now, and consequently the Waddingtonian view remained largely contained within the small world of developmental biology specialists. The discovery of DNA structure, and the accumulation of knowledge on molecular basis of gene regulation that followed, led to a shift in the understanding of epigenetics from a conceptual framework to understand development, as proposed by Waddington, to the study of proximal molecular determinants of gene expression, as subsequently used in practice.<sup>[6,26,59]</sup> The origin of this latter interpretation of epigenetics can be found in David L. Nanney's distinction between genetic systems and epigenetic systems, the latter defined as "auxiliary mechanisms with different principles of operation involved in determining which specificities are to be expressed in any particular cell".<sup>[6,60]</sup> Nanney took the term 'epigenetic' from Waddington; but instead of a general model of processes linking genotype and phenotype, he focused on the role cytoplasmic and extranuclear molecular factors have in heredity.<sup>[3,6,60]</sup> During the 1970s-80s, the discovery of the effects of DNA methylation and histone covalent modifications on gene expression provided concrete material proof of the epigenetic systems proposed by Nanney.<sup>[59,61]</sup> This variety of 'epigenetic' phenomena referred to as epigenetic stems from an etymological division of the word into the prefix "epi-", meaning above, and "genetic", referring to information encoded in the DNA sequence. Since then, advances in molecular biology and technological developments have expanded the number of known epigenetic chemical modifications influencing gene expression.<sup>[11]</sup>

## Box 2: Modelling the epigenetic landscape of attractors

Classical dynamical systems theory seeks to understand the way a system (a collection of interacting elements) will behave through time by analyzing it mathematically, as shown in Fig. 2b. At the center of the theoretical approach that concerns the study of dynamical systems is the concept of state, see Fig. 2c. This refers to a minimum set of descriptive variables of the system that correspond to a kind of memory that is constantly updated. Then, the system can be described at any point of time by the vector of states of its constituent elements. This means that, depending on the system's state and the regulatory rules among them, the behavior of the system is causally determined. From a very general point of view the system can stay in the same state it is or change to a different state. When a system reaches a state that does not change in time, this state is known as an attractor state and the system's regulatory logic will make it remain there. Some systems, known as multistable systems, have several attractor states. Going back to an organism, the set of attractor states of such a complex multidimensional system corresponds to the different cell types reached during development.

A dynamical system with two variables can be graphically represented by a Cartesian plane. This abstraction is known as the system's state space and contains all the possible states of the system. The system's dynamics can be represented as trajectories in the state space, which will eventually reach an attractor state. The state space can be expanded to dynamical systems of any number of dimensions, although we cannot perceive them in visual terms for systems bigger than three dimensions, generating  $n$  dimensional state spaces in which every spatial dimension corresponds to a system's variable, as depicted in Fig. 2d. The dynamical systems framework provides a conceptual way to formally take into account the possibility of the system to leave a given attractor and reach another one, corresponding to cell type transitions observed in development. Attractor transitions are analyzed by adding stochasticity to the regulatory network model, reflecting the stochastic nature of transcription, translation, and chemical reactions in the cell.<sup>[37,62]</sup> By introducing stochasticity to the model the probability to leave an attractor can be measured, as shown in Fig. 2e. This probability is known as the attractor's relative stability, and from Waddington's epigenetic landscape idea it is analogous to the inverse of the valley's depth.<sup>[31,62,63]</sup> In this way, the integration of stochastic noise to the regulatory network model results in a probabilistic landscape that captures the possibility of transitioning between attractors and relates directly to Waddington's diagram of the epigenetic landscape, see Fig. 2f.

## References

1. Bird, A. (2007). Perceptions of epigenetics. *Nature*, 447(7143), 396–398.
2. Deichmann, U. (2016). Epigenetics: The origins and evolution of a fashionable topic. In *Developmental Biology* (Vol. 416, Issue 1, pp. 249–254).  
<https://doi.org/10.1016/j.ydbio.2016.06.005>
3. Pisco, A. O., d'Hérouël, A. F., & Huang, S. (2016). Conceptual Confusion: The case of Epigenetics. In *BioRxiv*. <https://doi.org/10.1101/053009>
4. Bernstein, B. E., Meissner, A., & Lander, E. S. (2007). The Mammalian Epigenome. In *Cell* (Vol. 128, Issue 4, pp. 669–681). <https://doi.org/10.1016/j.cell.2007.01.033>
5. Rivera, C. M., & Ren, B. (2013). Mapping Human Epigenomes. In *Cell* (Vol. 155, Issue 1, pp. 39–55). <https://doi.org/10.1016/j.cell.2013.09.011>
6. Haig, D. (2004). The (Dual) Origin of Epigenetics. In *Cold Spring Harbor Symposia on Quantitative Biology* (Vol. 69, Issue 0, pp. 67–70).  
<https://doi.org/10.1101/sqb.2004.69.67>
7. Fagan, M. B. (2012). Waddington redux: models and explanation in stem cell and systems biology. In *Biology & Philosophy* (Vol. 27, Issue 2, pp. 179–213).  
<https://doi.org/10.1007/s10539-011-9294-y>
8. Natoli, G. (2010). Maintaining Cell Identity through Global Control of Genomic Organization. In *Immunity* (Vol. 33, Issue 1, pp. 12–24).  
<https://doi.org/10.1016/j.immuni.2010.07.006>
9. Roadmap Epigenomics Consortium, Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., Ziller, M. J., Amin, V., Whitaker, J. W., Schultz, M. D., Ward, L. D., Sarkar, A., Quon, G., Sandstrom, R. S., Eaton, M. L., ... Kellis, M. (2015). Integrative analysis of 111 reference human epigenomes. *Nature*, 518(7539), 317–330.
10. Goldberg, A. D., David Allis, C., & Bernstein, E. (2007). Epigenetics: A Landscape Takes Shape. In *Cell* (Vol. 128, Issue 4, pp. 635–638).

- 1  
2  
3 <https://doi.org/10.1016/j.cell.2007.02.006>  
4  
5  
6 11. Allis, C. D., David Allis, C., & Jenuwein, T. (2016). The molecular hallmarks of epigenetic  
7 control. In *Nature Reviews Genetics* (Vol. 17, Issue 8, pp. 487–500).  
8  
9 <https://doi.org/10.1038/nrg.2016.59>  
10  
11 12. Hemberger, M., Dean, W., & Reik, W. (2009). Epigenetic dynamics of stem cells and cell  
12 lineage commitment: digging Waddington's canal. In *Nature Reviews Molecular Cell*  
13 *Biology* (Vol. 10, Issue 8, pp. 526–537). <https://doi.org/10.1038/nrm2727>  
14  
15 13. Häfner, S. J., & Lund, A. H. (2016). Great expectations – Epigenetics and the  
16 meandering path from bench to bedside. In *Biomedical Journal* (Vol. 39, Issue 3, pp.  
17 166–176). <https://doi.org/10.1016/j.bj.2016.01.008>  
18  
19 14. Hawkins, R. D., Hon, G. C., Lee, L. K., Ngo, Q., Lister, R., Pelizzola, M., Edsall, L. E.,  
20 Kuan, S., Luu, Y., Klugman, S., Antosiewicz-Bourget, J., Ye, Z., Espinoza, C., Agarwahl,  
21 S., Shen, L., Ruotti, V., Wang, W., Stewart, R., Thomson, J. A., ... Ren, B. (2010).  
22 Distinct epigenomic landscapes of pluripotent and lineage-committed human cells. *Cell*  
23 *Stem Cell*, 6(5), 479–491.  
24  
25 15. Bernal, A. J., & Jirtle, R. L. (2010). Epigenomic disruption: The effects of early  
26 developmental exposures. In *Birth Defects Research Part A: Clinical and Molecular*  
27 *Teratology* (Vol. 88, Issue 10, pp. 938–944). <https://doi.org/10.1002/bdra.20685>  
28  
29 16. Marks, H., Kalkan, T., Menafrá, R., Denissov, S., Jones, K., Hofemeister, H., Nichols, J.,  
30 Kranz, A., Francis Stewart, A., Smith, A., & Stunnenberg, H. G. (2012). The  
31 Transcriptional and Epigenomic Foundations of Ground State Pluripotency. In *Cell* (Vol.  
32 149, Issue 3, pp. 590–604). <https://doi.org/10.1016/j.cell.2012.03.026>  
33  
34 17. Oyama, S. (2020). *The Ontogeny of Information*. <https://doi.org/10.1515/9780822380665>  
35  
36 18. Morgan, M. A. J., & Shilatifard, A. (2020). Reevaluating the roles of histone-modifying  
37 enzymes and their associated chromatin modifications in transcriptional regulation. In  
38 *Nature Genetics* (Vol. 52, Issue 12, pp. 1271–1281). [https://doi.org/10.1038/s41588-](https://doi.org/10.1038/s41588-020-00736-4)  
39 [020-00736-4](https://doi.org/10.1038/s41588-020-00736-4)  
40  
41 19. Trojer, P., & Reinberg, D. (2006). Histone Lysine Demethylases and Their Impact on  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

- 1  
2  
3 Epigenetics. In *Cell* (Vol. 125, Issue 2, pp. 213–217).  
4  
5 <https://doi.org/10.1016/j.cell.2006.04.003>  
6  
7 20. Chen, T., & Dent, S. Y. R. (2014). Chromatin modifiers and remodellers: regulators of  
8 cellular differentiation. In *Nature Reviews Genetics* (Vol. 15, Issue 2, pp. 93–106).  
9  
10 <https://doi.org/10.1038/nrg3607>  
11  
12 21. Lappalainen, T., & Grealley, J. M. (2017). Associating cellular epigenetic models with  
13 human phenotypes. In *Nature Reviews Genetics* (Vol. 18, Issue 7, pp. 441–451).  
14  
15 <https://doi.org/10.1038/nrg.2017.32>  
16  
17 22. Knaupp, A. S., Buckberry, S., Pflueger, J., Lim, S. M., Ford, E., Larcombe, M. R.,  
18 Rossello, F. J., de Mendoza, A., Alaei, S., Firas, J., Holmes, M. L., Nair, S. S., Clark, S.  
19 J., Nefzger, C. M., Lister, R., & Polo, J. M. (2017). Transient and Permanent  
20 Reconfiguration of Chromatin and Transcription Factor Occupancy Drive  
21 Reprogramming. In *Cell Stem Cell* (Vol. 21, Issue 6, pp. 834–845.e6).  
22  
23 <https://doi.org/10.1016/j.stem.2017.11.007>  
24  
25 23. Domcke, S., Bardet, A. F., Ginno, P. A., Hartl, D., Burger, L., & Schübeler, D. (2015).  
26 Competition between DNA methylation and transcription factors determines binding of  
27 NRF1. In *Nature* (Vol. 528, Issue 7583, pp. 575–579).  
28  
29 <https://doi.org/10.1038/nature16462>  
30  
31 24. Feldmann, A., Ivanek, R., Murr, R., Gaidatzis, D., Burger, L., & Schübeler, D. (2013).  
32 Transcription Factor Occupancy Can Mediate Active Turnover of DNA Methylation at  
33 Regulatory Regions. In *PLoS Genetics* (Vol. 9, Issue 12, p. e1003994).  
34  
35 <https://doi.org/10.1371/journal.pgen.1003994>  
36  
37 25. Zaret, K. S., & Mango, S. E. (2016). Pioneer transcription factors, chromatin dynamics,  
38 and cell fate control. *Current Opinion in Genetics & Development*, 37, 76–81.  
39  
40 26. Henikoff, S., & Grealley, J. M. (2016). Epigenetics, cellular memory and gene regulation.  
41  
42 In *Current Biology* (Vol. 26, Issue 14, pp. R644–R648).  
43  
44 <https://doi.org/10.1016/j.cub.2016.06.011>  
45  
46 27. Klemm, S. L., Shipony, Z., & Greenleaf, W. J. (2019). Chromatin accessibility and the  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

- 1  
2  
3 regulatory epigenome. In *Nature Reviews Genetics* (Vol. 20, Issue 4, pp. 207–220).  
4  
5 <https://doi.org/10.1038/s41576-018-0089-8>  
6  
7  
8 28. Villaseñor, R., & Baubec, T. (2021). Regulatory mechanisms governing chromatin  
9 organization and function. In *Current Opinion in Cell Biology* (Vol. 70, pp. 10–17).  
10  
11 <https://doi.org/10.1016/j.ceb.2020.10.015>  
12  
13  
14 29. Huang, S., & Kauffman, S. A. (2012). ComplexGRN complex GeneComplex GRN  
15 Regulatory Networks – from Structure to Biological Observables: Cell Fate  
16 DeterminationGene regulation, cell fate determination. In *Computational Complexity* (pp.  
17 527–560). [https://doi.org/10.1007/978-1-4614-1800-9\\_35](https://doi.org/10.1007/978-1-4614-1800-9_35)  
18  
19  
20  
21  
22 30. Waddington, C. H. (2012). The Epigenotype. In *International Journal of Epidemiology*  
23 (Vol. 41, Issue 1, pp. 10–13). <https://doi.org/10.1093/ije/dyr184>  
24  
25  
26  
27 31. Huang, S. (2012). The molecular and mathematical basis of Waddington’s epigenetic  
28 landscape: A framework for post-Darwinian biology? In *BioEssays* (Vol. 34, Issue 2, pp.  
29 149–157). <https://doi.org/10.1002/bies.201100031>  
30  
31  
32  
33 32. Kauffman, S. A. (1969). Metabolic stability and epigenesis in randomly constructed  
34 genetic nets. *Journal of Theoretical Biology*, 22(3), 437–467.  
35  
36  
37 33. Huang, S., Eichler, G., Bar-Yam, Y., & Ingber, D. E. (2005). Cell fates as high-  
38 dimensional attractor states of a complex gene regulatory network. *Physical Review*  
39 *Letters*, 94(12), 128701.  
40  
41  
42  
43 34. Davidson, E. H. (2010). *The Regulatory Genome: Gene Regulatory Networks In*  
44 *Development And Evolution*. Elsevier.  
45  
46  
47 35. Alvarezbuylla, E., Benitez, M., Davila, E., Chaos, A., Espinosasoto, C., & Padillalongoria,  
48 P. (2007). Gene regulatory network models for plant development. In *Current Opinion in*  
49 *Plant Biology* (Vol. 10, Issue 1, pp. 83–91). <https://doi.org/10.1016/j.pbi.2006.11.008>  
50  
51  
52  
53 36. Newman, S. A. (2020). Cell differentiation: What have we learned in 50 years? *Journal of*  
54 *Theoretical Biology*, 485, 110031.  
55  
56  
57  
58 37. Niklas, K. J., Bondos, S. E., Dunker, A. K., & Newman, S. A. (2015). Rethinking gene  
59 regulatory networks in light of alternative splicing, intrinsically disordered protein  
60



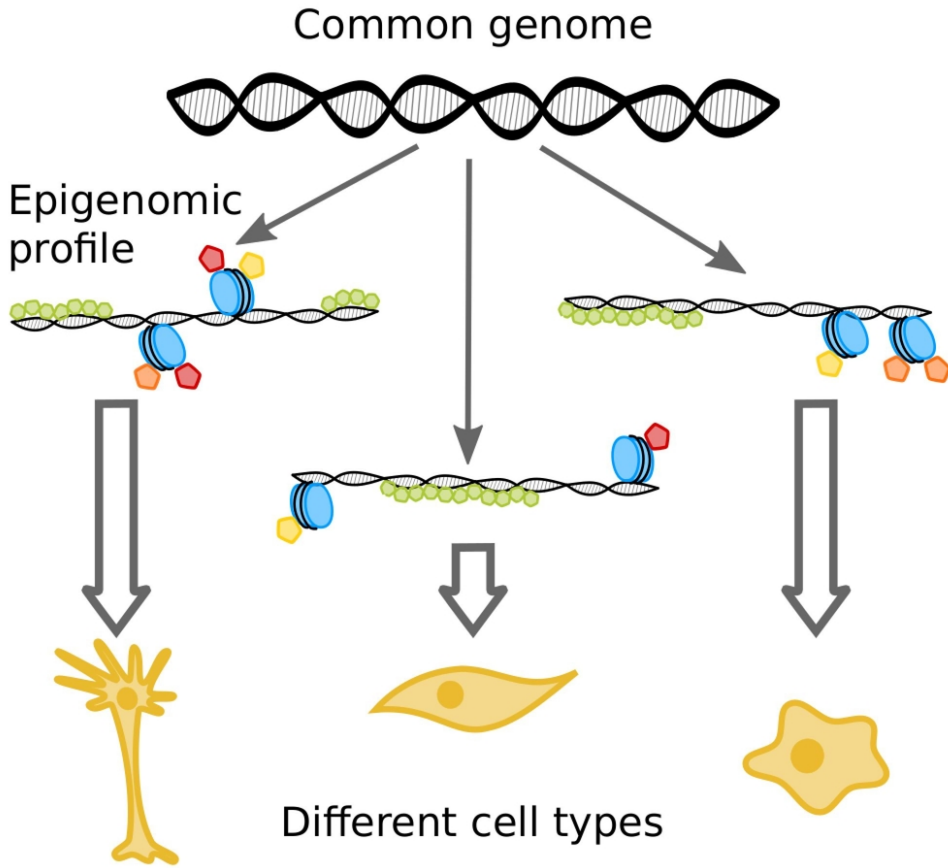
- domains, and post-translational modifications. *Frontiers in Cell and Developmental Biology*, 3, 8.
38. Staby, L., O'Shea, C., Willemoës, M., Theisen, F., Kragelund, B. B., & Skriver, K. (2017). Eukaryotic transcription factors: paradigms of protein intrinsic disorder. *Biochemical Journal*, 474(15), 2509–2532.
39. Prohaska, S. J., Stadler, P. F., & Krakauer, D. C. (2010). Innovation in gene regulation: the case of chromatin computation. *Journal of Theoretical Biology*, 265(1), 27–44.
40. Wang, J., Sun, H., Jiang, M., Li, J., Zhang, P., Chen, H., Mei, Y., Fei, L., Lai, S., Han, X., Song, X., Xu, S., Chen, M., Ouyang, H., Zhang, D., Yuan, G.-C., & Guo, G. (2021). Tracing cell-type evolution by cross-species comparison of cell atlases. In *Cell Reports* (Vol. 34, Issue 9, p. 108803). <https://doi.org/10.1016/j.celrep.2021.108803>
41. Sebé-Pedrós, A., Degnan, B. M., & Ruiz-Trillo, I. (2017). The origin of Metazoa: a unicellular perspective. In *Nature Reviews Genetics* (Vol. 18, Issue 8, pp. 498–512). <https://doi.org/10.1038/nrg.2017.21>
42. Sebé-Pedrós, A., Chomsky, E., Pang, K., Lara-Astiaso, D., Gaiti, F., Mukamel, Z., Amit, I., Hejnol, A., Degnan, B. M., & Tanay, A. (2018). Early metazoan cell type diversity and the evolution of multicellular gene regulation. In *Nature Ecology & Evolution* (Vol. 2, Issue 7, pp. 1176–1188). <https://doi.org/10.1038/s41559-018-0575-6>
43. Lewontin, R. (2020). Foreword. In *The Ontogeny of Information* (pp. vii – xvi). <https://doi.org/10.1515/9780822380665-001>
44. Neph, S., Stergachis, A. B., Reynolds, A., Sandstrom, R., Borenstein, E., & Stamatoyannopoulos, J. A. (2012). Circuitry and Dynamics of Human Transcription Factor Regulatory Networks. In *Cell* (Vol. 150, Issue 6, pp. 1274–1286). <https://doi.org/10.1016/j.cell.2012.04.040>
45. Stergachis, A. B., Neph, S., Sandstrom, R., Haugen, E., Reynolds, A. P., Zhang, M., Byron, R., Canfield, T., Stelhing-Sun, S., Lee, K., Thurman, R. E., Vong, S., Bates, D., Neri, F., Diegel, M., Giste, E., Dunn, D., Vierstra, J., Scott Hansen, R., ... Stamatoyannopoulos, J. A. (2014). Conservation of trans-acting circuitry during

- 1  
2  
3 mammalian regulatory evolution. In *Nature* (Vol. 515, Issue 7527, pp. 365–370).  
4  
5 <https://doi.org/10.1038/nature13972>  
6  
7  
8 46. Marbach, D., Lamparter, D., Quon, G., Kellis, M., Kutalik, Z., & Bergmann, S. (2016).  
9 Tissue-specific regulatory circuits reveal variable modular perturbations across complex  
10 diseases. *Nature Methods*, 13(4), 366–370.  
11  
12  
13 47. Cantone, I., & Fisher, A. G. (2013). Epigenetic programming and reprogramming during  
14 development. In *Nature Structural & Molecular Biology* (Vol. 20, Issue 3, pp. 282–289).  
15  
16 <https://doi.org/10.1038/nsmb.2489>  
17  
18  
19 48. Lindroth, A. M., Park, Y. J., & Plass, C. (2015). Epigenetic Reprogramming in Cancer. In  
20 *Epigenetic Mechanisms in Cellular Reprogramming* (pp. 193–223).  
21  
22 [https://doi.org/10.1007/978-3-642-31974-7\\_9](https://doi.org/10.1007/978-3-642-31974-7_9)  
23  
24  
25 49. Ladstätter, S., & Tachibana, K. (2019). Genomic insights into chromatin reprogramming  
26 to totipotency in embryos. In *Journal of Cell Biology* (Vol. 218, Issue 1, pp. 70–82).  
27  
28 <https://doi.org/10.1083/jcb.201807044>  
29  
30  
31 50. Seisenberger, S., Peat, J. R., Hore, T. A., Santos, F., Dean, W., & Reik, W. (2013).  
32 Reprogramming DNA methylation in the mammalian life cycle: building and breaking  
33 epigenetic barriers. *Philosophical Transactions of the Royal Society of London. Series*  
34 *B, Biological Sciences*, 368(1609), 20110330.  
35  
36  
37 51. Gao, L., Wu, K., Liu, Z., Yao, X., Yuan, S., Tao, W., Yi, L., Yu, G., Hou, Z., Fan, D., Tian,  
38 Y., Liu, J., Chen, Z.-J., & Liu, J. (2018). Chromatin Accessibility Landscape in Human  
39 Early Embryos and Its Association with Evolution. In *Cell* (Vol. 173, Issue 1, pp. 248–  
40 259.e15). <https://doi.org/10.1016/j.cell.2018.02.028>  
41  
42  
43 52. Shahbazi, M. N., Siggia, E. D., & Zernicka-Goetz, M. (2019). Self-organization of stem  
44 cells into embryos: A window on early mammalian development. In *Science* (Vol. 364,  
45 Issue 6444, pp. 948–951). <https://doi.org/10.1126/science.aax0164>  
46  
47  
48 53. Zhu, M., & Zernicka-Goetz, M. (2020). Principles of Self-Organization of the Mammalian  
49 Embryo. In *Cell* (Vol. 183, Issue 6, pp. 1467–1478).  
50  
51  
52 <https://doi.org/10.1016/j.cell.2020.11.003>  
53  
54  
55  
56  
57  
58  
59  
60

- 1  
2  
3 54. Fagan, M. B. (2013). Philosophy of Stem Cell Biology - an Introduction. In *Philosophy*  
4 *Compass* (Vol. 8, Issue 12, pp. 1147–1158). <https://doi.org/10.1111/phc3.12088>  
5  
6  
7 55. Meshorer, E., & Misteli, T. (2006). Chromatin in pluripotent embryonic stem cells and  
8 differentiation. In *Nature Reviews Molecular Cell Biology* (Vol. 7, Issue 7, pp. 540–546).  
9 <https://doi.org/10.1038/nrm1938>  
10  
11  
12  
13 56. Boskovic, A., Eid, A., Pontabry, J., Ishiuchi, T., Spiegelhalter, C., Raghu Ram, E. V.,  
14 Meshorer, E., & Torres-Padilla, M.-E. (2014). Higher chromatin mobility supports  
15 totipotency and precedes pluripotency in vivo. In *Genes & Development* (Vol. 28, Issue  
16 10, pp. 1042–1047). <https://doi.org/10.1101/gad.238881.114>  
17  
18  
19  
20  
21 57. Lim, P. S. L., & Meshorer, E. (2021). Organization of the Pluripotent Genome. In *Cold*  
22 *Spring Harbor Perspectives in Biology* (Vol. 13, Issue 2, p. a040204).  
23 <https://doi.org/10.1101/cshperspect.a040204>  
24  
25  
26  
27 58. Waddington, C. H. (2008). The Basic Ideas of Biology. In *Biological Theory* (Vol. 3, Issue  
28 3, pp. 238–253). <https://doi.org/10.1162/biot.2008.3.3.238>  
29  
30  
31  
32 59. Felsenfeld, G. (2014). A brief history of epigenetics. *Cold Spring Harbor Perspectives in*  
33 *Biology*, 6(1). <https://doi.org/10.1101/cshperspect.a018200>  
34  
35  
36 60. Nanney, D. L. (1958). EPIGENETIC CONTROL SYSTEMS. In *Proceedings of the*  
37 *National Academy of Sciences* (Vol. 44, Issue 7, pp. 712–717).  
38 <https://doi.org/10.1073/pnas.44.7.712>  
39  
40  
41  
42 61. Holliday, R. (2006). Epigenetics: a historical overview. *Epigenetics: Official Journal of the*  
43 *DNA Methylation Society*, 1(2), 76–80.  
44  
45  
46 62. Mojtahedi, M., Skupin, A., Zhou, J., Castaño, I. G., Leong-Quong, R. Y. Y., Chang, H.,  
47 Trachana, K., Giuliani, A., & Huang, S. (2016). Cell Fate Decision as High-Dimensional  
48 Critical State Transition. In *PLOS Biology* (Vol. 14, Issue 12, p. e2000640).  
49 <https://doi.org/10.1371/journal.pbio.2000640>  
50  
51  
52  
53 63. Davila-Velderrain, J., Martinez-Garcia, J. C., & Alvarez-Buylla, E. R. (2015). Modeling  
54 the epigenetic attractors landscape: toward a post-genomic mechanistic understanding  
55 of development. In *Frontiers in Genetics* (Vol. 6).  
56  
57  
58  
59  
60

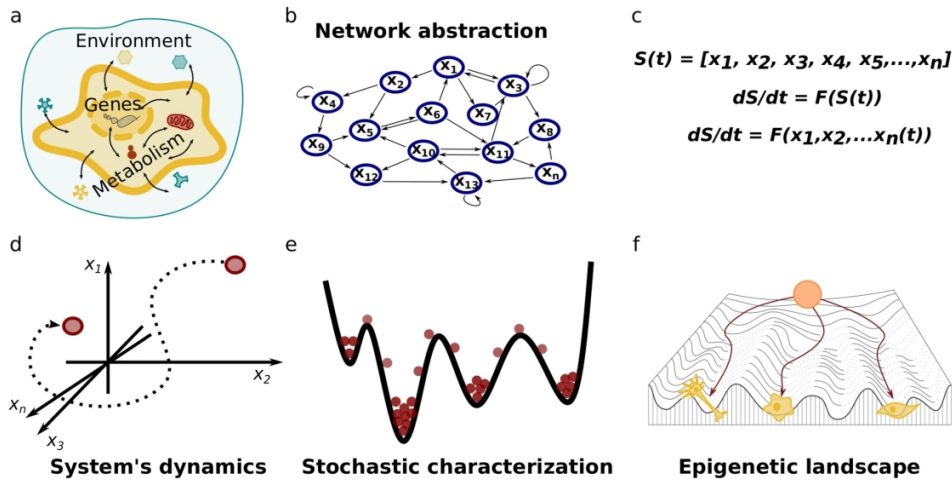
1  
2  
3 <https://doi.org/10.3389/fgene.2015.00160>  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

For Peer Review



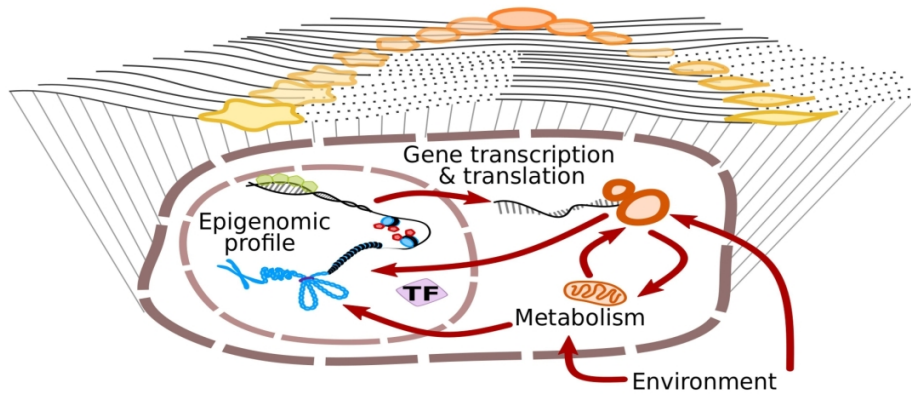
Molecular epigenomics grants epigenomic marks an explanatory condition for cell phenotypic differences. The phenotypical differences among cell types of an organism are explained by the differences in the configuration of their epigenome.

85x80mm (300 x 300 DPI)



Dynamical systems modeling of the epigenetic landscape. The epigenetic landscape framework conceives the cell as a) a multidimensional system in which the molecules involved in its behavior are represented as b) a network of interacting elements. c) The dynamical behavior of this network is analyzed by a mathematical formalization. d) The complete set of possible system states defines the state space, where each state corresponds to a specific system configuration. Thus, developmental dynamics are described as trajectories in this abstract space. e) Introducing stochasticity to the model grants the characterization of the system states in terms of relative stability, and f) the intuitive interpretation of the associated epigenetic landscape.

170x90mm (300 x 300 DPI)



Integrating epigenomics to the epigenetic landscape. To understand the emergency and existence of different cell types during development; epigenomic marks should be integrated to the network of regulatory elements of the cell, their presence resulting from the regulatory interactions of the underlying system and influencing its dynamics.

170x70mm (300 x 300 DPI)

# Bibliografía

- [1] Garcia-Ojalvo, J. and Martinez Arias, A. (2012) Towards a statistical mechanics of cell fate decisions. *Current Opinion in Genetics and Development development*, **22**(6), 619–626.
- [2] Fagan, M. B. (2013) Philosophy of stem cell biology - An introduction. *Philosophy Compass*, **8**(12), 1147–1158.
- [3] De Los Angeles, A., Ferrari, F., Xi, R., Fujiwara, Y., Benvenisty, N., Deng, H., Hochedlinger, K., Jaenisch, R., Lee, S., Leitch, H. G., Lensch, M. W., Lujan, E., Pei, D., Rossant, J., Wernig, M., Park, P. J., and Daley, G. Q. (2015) Hallmarks of pluripotency. *Nature*, **525**(7570), 469–78.
- [4] Gilbert, S. F. and Barresi, M. J. F. (2018) *Developmental Biology*, Sinauer Associates, Inc., Massachusetts, U.S.A. 11th ed. edition.
- [5] Mazzarello, P. (1999) A unifying concept: the history of cell theory. *Nature cell biology*, **1**(1), E13–E15.
- [6] Harris, H. (2000) *The birth of the cell*, Yale University Press, .
- [7] Amit, I., Bader, G., Campbell, P., Carninci, P., Clevers, H., Eils, R., Hacohen, N., Kriegstein, A., Lander, E., Linnarsson, S., Majumdar, P., Merad, M., Naik, S., Nolan, G., Pe'er, D., Ponting, C., Quake, S., Rajewsky, N., Regev, A., Shapiro, E., Shin, J., Stratton, M., Stunnenberg, H., Teichmann, S., van Oudenaarden, A., Weissman, J., and Wold, B. (2018) The human cell atlas. *arXiv*, pp. 1–31.
- [8] Martinez Arias, A. and Brickman, J. M. (2011) Gene expression heterogeneities in embryonic stem cell populations: Origin and function. *Current Opinion in Cell Biology*, **23**(6), 650–656.
- [9] Regev, A. and Teichmann, S. A. (2017) The Human Cell Atlas: from vision to reality. *Nature*, **550**.
- [10] Snippert, H. J. and Clevers, H. (2011) Tracking adult stem cells. *EMBO reports*, **12**(2), 113–122.
- [11] Baker, C. L. and Pera, M. F. (2018) Capturing Totipotent Stem Cells. *Cell Stem Cell*, **22**(1), 25–34.
- [12] Martello, G. and Smith, A. (2014) The nature of embryonic stem cells. *Annual review of cell and developmental biology*, **30**, 647–675.



- [13] Zipori, D. (2004) The nature of stem cells: state rather than entity. *Nature Reviews Genetics*, **5**(11), 873–878.
- [14] Ivanova, N. B., Dimos, J. T., Schaniel, C., Hackney, J. A., Moore, K. A., and Lemischka, I. R. (2002) A stem cell molecular signature. *Science*, **298**(5593), 601–604.
- [15] Ramalho-Santos, M., Yoon, S., Matsuzaki, Y., Mulligan, R. C., and Melton, D. A. (2002) "Stemness": transcriptional profiling of embryonic and adult stem cells. *Science*, **298**(5593), 597–600.
- [16] Fortunel, N. O., Otu, H. H., Ng, H.-H., and Chen, J. (2003) Comment on “‘Stemness’: Transcriptional Profiling of Embryonic and Adult Stem Cells” and “A Stem Cell Molecular signature” (I). *Science*, **302**(5644), 393b.
- [17] Vogel, G. (2003) ‘ Stemness ’ Genes Still Elusive. *Science*, **302**, 371.
- [18] Herberg, M. and Roeder, I. (2015) Computational modelling of embryonic stem-cell fate control.. *Development (Cambridge, England)*, **142**(13), 2250–60.
- [19] Adler, C. E. and Sánchez Alvarado, A. (2015) Types or States? Cellular Dynamics and Regenerative Potential. *Trends in Cell Biology*, **25**(11), 687–696.
- [20] Solter, D. (2006) From teratocarcinomas to embryonic stem cells and beyond: a history of embryonic stem cell research. *Nature Reviews Genetics*, **7**(4), 319–327.
- [21] Kaneko, K. (2011) Characterization of stem cells and cancer cells on the basis of gene expression profile stability, plasticity, and robustness: Dynamical systems theory of gene expressions under cell-cell interaction explains mutational robustness of differentiated cells . *BioEssays*, **33**(6), 403–413.
- [22] Torres-Padilla, M. E. and Chambers, I. (2014) Transcription factor heterogeneity in pluripotent stem cells: A stochastic advantage. *Development (Cambridge)*, **141**(11), 2173–2181.
- [23] Roeder, I. and Radtke, F. (2009) Stem cell biology meets systems biology. *Development (Cambridge, England)*, **136**(21), 3525–3530.
- [24] Wen, L. and Tang, F. (2016) Single-cell sequencing in stem cell biology. *Genome Biology*, **17**(1), 1–12.
- [25] Kumar, P., Tan, Y., and Cahan, P. (2017) Understanding development and stem cells using single cell-based analyses of gene expression. *Development (Cambridge)*, **144**(1), 17–32.
- [26] Gulati, G. S., Sikandar, S. S., Wesche, D. J., Manjunath, A., Bharadwaj, A., Berger, M. J., Ilagan, F., Kuo, A. H., Hsieh, R. W., Cai, S., Zabala, M., Scheeren, F. A., Lobo, N. A., Qian, D., Yu, F. B., Dirbas, F. M., Clarke, M. F., and Newman, A. M. (2020) Single-cell transcriptional diversity is a hallmark of developmental potential. *Science*, **367**(6476), 405–411.

- [27] MacArthur, B. D. and Lemischka, I. R. (2013) Statistical mechanics of pluripotency. *Cell*, **154**(3), 484–489.
- [28] Huang, S. (2009) Non-genetic heterogeneity of cells in development: more than just noise. *Development*, **136**(23), 3853–3862.
- [29] Dueck, H., Eberwine, J., and Kim, J. (2016) Variation is function: Are single cell differences functionally important?: Testing the hypothesis that single cell variation is required for aggregate function. *BioEssays*, **38**(2), 172–180.
- [30] Moris, N., Pina, C., and Martinez Arias, A. (2016) Transition states and cell fate decisions in epigenetic landscapes. *Nature Reviews Genetics*, **17**(11), 693–703.
- [31] Goldberg, A. D., Allis, C. D., and Bernstein, E. (2007) Epigenetics: A Landscape Takes Shape. *Cell*, **128**(4), 635–638.
- [32] Allis, C. D. and Jenuwein, T. (2016) The molecular hallmarks of epigenetic control. *Nature Reviews Genetics*, **17**(8), 487–500.
- [33] Meshorer, E. and Misteli, T. (2006) Chromatin in pluripotent embryonic stem cells and differentiation. *Nature Reviews Molecular Cell Biology*, **7**(7), 540–546.
- [34] Gaspar-Maia, A., Alajem, A., Meshorer, E., and Ramalho-Santos, M. (2011) Open chromatin in pluripotency and reprogramming. *Nature Reviews Molecular Cell Biology*, **12**(1), 36–47.
- [35] Dixon, J. R., Jung, I., Selvaraj, S., Shen, Y., Antosiewicz-Bourget, J. E., Lee, A. Y., Ye, Z., Kim, A., Rajagopal, N., Xie, W., Diao, Y., Liang, J., Zhao, H., Lobanenko, V. V., Ecker, J. R., Thomson, J. a., and Ren, B. (2015) Chromatin architecture reorganization during stem cell differentiation. *Nature*, **518**(7539), 331–336.
- [36] Freire-Pritchett, P., Schoenfelder, S., Várnai, C., Wingett, S. W., Cairns, J., Collier, A. J., García-Vílchez, R., Furlan-Magaril, M., Osborne, C. S., Fraser, P., Rugg-Gunn, P. J., and Spivakov, M. (mar, 2017) Global reorganisation of cis -regulatory units upon lineage commitment of human embryonic stem cells. *eLife*, **6**, 1–26.
- [37] Lim, P. S. and Meshorer, E. (2020) Organization of the Pluripotent Genome. *Cold Spring Harbor Perspectives in Biology*, p. a040204.
- [38] Harikumar, A. and Meshorer, E. (2015) Chromatin remodeling and bivalent histone modifications in embryonic stem cells. *EMBO reports*, **16**(12), 1609–1619.
- [39] Blanco, E., González-Ramírez, M., Alcaine-Colet, A., Aranda, S., and Di Croce, L. (2020) The bivalent genome: characterization, structure, and regulation. *Trends in Genetics*, **36**(2), 118–131.

- [40] Hemberger, M., Dean, W., and Reik, W. (2009) Epigenetic dynamics of stem cells and cell lineage commitment: Digging Waddington's canal. *Nature Reviews Molecular Cell Biology*, **10**(8), 526–537.
- [41] Lander, A. D. (2010) The edges of understanding.. *BMC biology*, **8**, 40.
- [42] Waddington, C. H. (1957) *The Strategy of the Genes. A Discussion of Some Aspects of Theoretical Biology*, Vol. 20, George Allen & Unwin Ltd., New York, USA.
- [43] Huang, S. (2004) Back to the biology in systems biology: What can we learn from biomolecular networks?. *Briefings in Functional Genomics and Proteomics*, **2**(4), 279–297.
- [44] Davila-Velderrain, J. and Alvarez-Buylla, E. R. (2015) Modeling the epigenetic attractors landscape: toward a post-genomic mechanistic understanding of development. *Frontiers in Genetics*, **6**(April), 1–14.
- [45] Huang, S., Eichler, G., Bar-yam, Y., and Ingber, D. E. (2005) Cell Fates as High-Dimensional Attractor States of a Complex Gene Regulatory Network. *Physical Review Letters*, **128701**(April), 1–4.
- [46] Huang, S. and Kauffman, S. A. (2012) Complex Gene Regulatory Networks – from Structure to Biological Observables: Cell Fate Determination. *Computational Complexity: Theory, Techniques, and Applications*, pp. 527–560.
- [47] Oliveira Pisco, A., Fouquier d'Herouel, A., and Huang, S. (2016) Conceptual confusion: The case of epigenetics. *bioRxiv*.
- [48] Newman, S. A. (jan, 2020) Cell differentiation: What have we learned in 50 years?. *Journal of Theoretical Biology*, **485**(July), 110031.
- [49] Kauffman, S. A. (1969) Metabolic stability and epigenesis in randomly constructed genetic nets.. *Journal of theoretical biology*, **22**(3), 437–467.
- [50] Davidson, E. H. (2010) *The regulatory genome: gene regulatory networks in development and evolution*, Elsevier, .
- [51] Alvarez-Buylla, E. R., Benítez, M., Balleza Davila, E., Chaos Cador, Á. C., Espinosa-Soto, C., Padilla-Longoria, P., and et al. (2007) Gene regulatory network models for plant development. *Current Opinion in Plant Biology*, **10**(1), 83–91.
- [52] Kaneko, K. and Yomo, T. (1994) Cell division, differentiation, and dynamic clustering. *Physica D*, **75**(97), 89–102.
- [53] Furusawa, C. and Kaneko, K. (2009) Chaotic expression dynamics implies pluripotency: When theory and experiment meet. *Biology Direct*, **4**.
- [54] Furusawa, C. and Kaneko, K. (2012) A Dynamical-Systems View of Stem Cell Biology. *Science*, **338**(6104), 215–217.

- [55] 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 of London. Series B, Biological sciences*, **366**(1575), 2247–2259.
- [56] Dunn, S.-J., Martello, G., Yordanov, B., Emmott, S., and Smith, a. G. (jun, 2014) Defining an essential transcription factor program for naïve pluripotency.. *Science (New York, N.Y.)*, **344**(6188), 1156–60.
- [57] Miyamoto, T., Furusawa, C., and Kaneko, K. (2015) Pluripotency, Differentiation, and Reprogramming: A Gene Expression Dynamics Model with Epigenetic Feedback Regulation. *PLoS Computational Biology*, **11**(8).
- [58] Huang, S. (2009) Reprogramming cell fates: Reconciling rarity with robustness. *BioEssays*, **31**(5), 546–560.
- [59] Boyer, L. A., Lee, T. I., Cole, M. F., Johnstone, S. E., Levine, S. S., Zucker, J. P., Guenther, M. G., Kumar, R. M., Murray, H. L., Jenner, R. G., et al. (2005) Core transcriptional regulatory circuitry in human embryonic stem cells. *cell*, **122**(6), 947–956.
- [60] Kim, J., Chu, J., Shen, X., Wang, J., and Orkin, S. H. (2008) An Extended Transcriptional Network for Pluripotency of Embryonic Stem Cells. *Cell*, **132**(6), 1049–1061.
- [61] Benitah, S. A., Bracken, A., Dou, Y., Huangfu, D., Ivanova, N., Koseki, H., Laurent, L., Lim, D. A., Meshorer, E., Pombo, A., Sander, M., and Xu, G. L. (2014) Stem cell epigenetics: Looking forward. *Cell Stem Cell*, **14**(6), 706.
- [62] Gifford, C. A., Ziller, M. J., Gu, H., Trapnell, C., Donaghey, J., Tsankov, A., Shalek, A. K., Kelley, D. R., Shishkin, A. A., Issner, R., Zhang, X., Coyne, M., Fostel, J. L., Holmes, L., Meldrim, J., Guttman, M., Epstein, C., Park, H., Kohlbacher, O., Rinn, J., Gnirke, A., Lander, E. S., Bernstein, B. E., and Meissner, A. (2013) Transcriptional and epigenetic dynamics during specification of human embryonic stem cells. *Cell*, **153**(5), 1149–1163.
- [63] Stergachis, A. B., Neph, S. J., Reynolds, A., Humbert, R., Miller, B., Paige, S. L., Vernot, B., Cheng, J. B., Thurman, R. E., Sandstrom, R., Haugen, E., Heimfeld, S., Murry, C. E., Akey, J. M., and Stamatoyannopoulos, J. A. (2013) Developmental fate and cellular maturity encoded in human regulatory DNA landscapes. *Cell*, **154**(4), 888–903.
- [64] Prohaska, S. J., Stadler, P. F., and Krakauer, D. C. (2010) Innovation in gene regulation: The case of chromatin computation. *Journal of Theoretical Biology*, **265**(1), 27–44.
- [65] Chen, T. and Dent, S. Y. R. (2014) Chromatin modifiers and remodellers: regulators of cellular differentiation.. *Nature reviews. Genetics*, **15**(2), 93–106.
- [66] Forgacs, G. and Newman, S. A. (2005) Biological physics of the developing embryo, Cambridge University Press, .

- [67] Huang, S. (2012) The molecular and mathematical basis of Waddington's epigenetic landscape: A framework for post-Darwinian biology?. *BioEssays*, **34**(2), 149–157.
- [68] Kauffman, S. A. and Strohman, R. C. (1994) *The Origins of Order: self organization and selection in evolution*, Vol. 454, Oxford university press New York, .
- [69] Oyama, S. (2000) *The ontogeny of information: Developmental systems and evolution*, Duke university press, .
- [70] Laubenbacher, R., Hower, V., Jarrah, A., Torti, S. V., Mendes, P., Torti, F. M., and Akman, S. (2010) A Systems Biology View of Cancer. *Biochim Biophys Acta*, **1796**(2), 129–139.
- [71] Soto, A. M. and Sonnenschein, C. (2004) The somatic mutation theory of cancer: growing problems with the paradigm?. *Bioessays*, **26**(10), 1097–1107.
- [72] Huang, S. (2011) On the intrinsic inevitability of cancer: From foetal to fatal attraction. *Seminars in Cancer Biology*, **21**(3), 183–199.
- [73] Li, S., Zhu, X., Liu, B., Wang, G., and Ao, P. (2015) Endogenous molecular network reveals two mechanisms of heterogeneity within gastric cancer. *Oncotarget*, **6**(15), 13607.
- [74] Yuan, R., Zhu, X., Radich, J. P., and Ao, P. (2016) From molecular interaction to acute promyelocytic leukemia: Calculating leukemogenesis and remission from endogenous molecular-cellular network. *Scientific reports*, **6**(1), 1–11.
- [75] Méndez-López, L. F., Davila-Velderrain, J., Domínguez-Hüttinger, E., Enríquez-Olguín, C., Martínez-García, J. C., and Alvarez-Buylla, E. R. (2017) Gene regulatory network underlying the immortalization of epithelial cells. *BMC systems biology*, **11**(1), 1–15.
- [76] Steinway, S. N., Zañudo, J. G., Ding, W., Rountree, C. B., Feith, D. J., Loughran, T. P., and Albert, R. (2014) Network modeling of TGF $\beta$  signaling in hepatocellular carcinoma epithelial-to-mesenchymal transition reveals joint sonic hedgehog and Wnt pathway activation. *Cancer research*, **74**(21), 5963–5977.
- [77] Aponte, P. M. and Caicedo, A. (2017) Stemness in cancer: Stem cells, cancer stem cells, and their microenvironment. *Stem Cells International*, **2017**.
- [78] Huang, S., Ernberg, I., and Kauffman, S. (2009) Cancer attractors: a systems view of tumors from a gene network dynamics and developmental perspective. In *Seminars in cell & developmental biology* Elsevier Vol. 20, pp. 869–876.
- [79] Sonnenschein, C. and Soto, A. M. (2018) Cancer-causing somatic mutations: they are neither necessary nor sufficient. *Organisms. Journal of Biological Sciences*, **2**(1), 55–62.
- [80] Martincorena, I. and Campbell, P. J. (2015) Somatic mutation in cancer and normal cells. *Science*, **349**(6255), 1483–1489.

- [81] Martincorena, I., Roshan, A., Gerstung, M., Ellis, P., Van Loo, P., McLaren, S., Wedge, D. C., Fullam, A., Alexandrov, L. B., Tubio, J. M., Stebbings, L., Menzies, A., Widaa, S., Stratton, M. R., Jones, P. H., and Campbell, P. J. (2015) High burden and pervasive positive selection of somatic mutations in normal human skin. *Science*, **348**(6237), 880–886.
- [82] Li, R., Di, L., Li, J., Fan, W., Liu, Y., Guo, W., Liu, W., Liu, L., Li, Q., Chen, L., Chen, Y., Miao, C., Liu, H., Wang, Y., Ma, Y., Xu, D., Lin, D., Huang, Y., Wang, J., Bai, F., and Wu, C. (2021) A body map of somatic mutagenesis in morphologically normal human tissues. *Nature*, **597**(7876), 398–403.
- [83] Kauffman, S. (1971) Differentiation of malignant to benign cells. *Journal of Theoretical Biology*, **31**(3), 429–451.
- [84] Barabási, A. L. and Oltvai, Z. N. (feb, 2004) Network biology: understanding the cell’s functional organization.. *Nature reviews. Genetics*, **5**(2), 101–13.
- [85] Marbach, D., Lamparter, D., Quon, G., Kellis, M., Kutalik, Z., and Bergmann, S. (2016) Tissue-specific regulatory circuits reveal variable modular perturbations across complex diseases.. *Nature methods*, pp. 1–44.
- [86] Neph, S., Stergachis, A. B., Reynolds, A., Sandstrom, R., Borenstein, E., and Stamatoyannopoulos, J. A. (sep, 2012) Circuitry and Dynamics of Human Transcription Factor Regulatory Networks. *Cell*, **150**(6), 1274–1286.
- [87] Stergachis, A. B., Neph, S., Sandstrom, R., Haugen, E., Reynolds, A. P., Zhang, M., Byron, R., Canfield, T., Stelting-Sun, S., Lee, K., Thurman, R. E., Vong, S., Bates, D., Neri, F., Diegel, M., Giste, E., Dunn, D., Vierstra, J., Hansen, R. S., Johnson, A. K., Sabo, P. J., Wilken, M. S., Reh, T. a., Treuting, P. M., Kaul, R., Groudine, M., Bender, M. a., Borenstein, E., and Stamatoyannopoulos, J. a. (2014) Conservation of trans-acting circuitry during mammalian regulatory evolution. *Nature*, **515**(7527), 365–370.
- [88] Albert, R., Jeong, H., and Barabási, A.-L. (jul, 2000) Error and attack tolerance of complex networks. *Nature*, **406**(6794), 378–382.
- [89] Lewontin, R. (2000) Foreword. In Oyama, S., (ed.), *The Ontogeny of Information*, chapter Foreword, pp. vii–xv Duke University Press 2nd edition.
- [90] Jaeger, J. and Monk, N. (sep, 2015) Everything flows: A process perspective on life. *EMBO reports*, **16**(9), 1064–107.
- [91] Consortium, E. P. et al. (2004) The ENCODE (ENCyclopedia of DNA elements) project. *Science*, **306**(5696), 636–640.
- [92] Bartha, I., di Iulio, J., Venter, J. C., and Telenti, A. (2018) Human gene essentiality. *Nature Reviews Genetics*, **19**(1), 51.
- [93] Pavličev, M. and Cheverud, J. M. (2015) Constraints evolve: context dependency of gene effects allows evolution of pleiotropy. *Annual Review of Ecology, Evolution, and Systematics*, **46**, 413–434.

- [94] Deichmann, U. (2016) Epigenetics: The origins and evolution of a fashionable topic. *Developmental Biology*, **416**(1), 249–254.
- [95] Grealley, J. M. (2018) A user’s guide to the ambiguous word ‘epigenetics’. *Nature Reviews Molecular Cell Biology*,.
- [96] Niklas, K. J., Bondos, S. E., Dunker, A. K., and Newman, S. A. (2015) Rethinking gene regulatory networks in light of alternative splicing, intrinsically disordered protein domains, and post-translational modifications. *Frontiers in Cell and Developmental Biology*, **3**(February), 1–13.
- [97] Margulis, L. and Sagan, D. (2000) What is life?, Univ of California Press, .