



UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

Maestría y Doctorado en Ciencias Bioquímicas

El interactoma en la red metabólica de *Saccharomyces cerevisiae*.
Análisis de centralidad y evaluación funcional de las enzimas indispensables para el crecimiento.

TESIS

QUE PARA OPTAR POR EL GRADO DE:
Maestro en Ciencias

PRESENTA:
Raúl Antonio Ortiz Merino

TUTOR PRINCIPAL
Dr. Gabriel del Río Guerra
INSTITUTO DE FISIOLÓGÍA CELULAR

MIEMBROS DEL COMITÉ TUTOR
Dra. Alicia González Manjarrez
INSTITUTO DE FISIOLÓGÍA CELULAR
Dr. Luis Antonio Mendoza Sierra
INSTITUTO DE INVESTIGACIONES BIOMÉDICAS

MÉXICO, D. F. abril, 2013



Universidad Nacional
Autónoma de México

Dirección General de Bibliotecas de la UNAM

Biblioteca Central



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.

Of. No. PMDCB/964/2012

RAÚL ANTONIO ORTIZ MERINO
Alumno (a) de la Maestría en Ciencias Bioquímicas
P r e s e n t e

Los miembros del Subcomité Académico, en reunión ordinaria del día 5 de Noviembre del presente año, conocieron su solicitud de ASIGNACIÓN de JURADO DE EXAMEN para optar por el grado de MAESTRO EN CIENCIAS (BIOQUIMICA), con la tesis titulada “El interactoma en la red metabólica de *Saccharomyces cerevisiae*. Análisis de centralidad y evaluación funcional de las enzimas indispensables para el crecimiento”, dirigida por el Dr. Gabriel del Río Guerra.

De su análisis se acordó ratificar al jurado asignado:

PRESIDENTE	Dra. Alejandra Covarrubias Robles
VOCAL	Dr. León Patricio Martínez Castilla
VOCAL	Dr. Juan Enrique Morett Sánchez
VOCAL	Dr. Lorenzo Patrick Segovia Forcella
SECRETARIO	Dr. Osbaldo Resendis Antonio

Sin otro particular por el momento, aprovecho la ocasión para enviarle un cordial saludo.

Atentamente
“POR MI RAZA HABLARÁ EL ESPÍRITU”
Cd. Universitaria, D.F., a 8 de Noviembre de 2012.
EL COORDINADOR DE ENTIDAD


DR. ROGELIO RODRÍGUEZ SOTRES

C.c.p. Archivo

RRS*lgg



mdcbq@posgrado.unam.mx

Of. No. PMDCBQ/416/2013

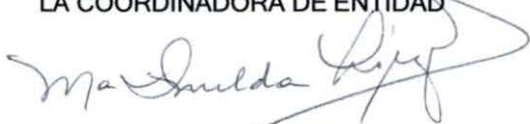
LIBB. RAÚL ANTONIO ORTIZ MERINO
Alumno de la Maestría en Ciencias Bioquímicas
P r e s e n t e

En base a su oficio enviado para la solicitud de cambio de integrantes de jurado de examen, el Subcomité Académico en la sesión ordinaria del día 6 de Mayo del presente acordó:

"Aceptar el cambio de sinodal sustituyendo al Dr. Osbaldo Resendis Antonio (Secretario) por la Dra. Emma C. Saavedra Lira"

Sin otro particular por el momento, aprovecho la ocasión para enviarle un cordial saludo.

Atentamente
"POR MI RAZA HABLARÁ EL ESPÍRITU"
Cd. Universitaria, D.F., a 13 de Mayo de 2013.
LA COORDINADORA DE ENTIDAD



DRA. MARÍA IMELDA LÓPEZ VILLASEÑOR

C.c.p. Archivo

MILV*lgg



**A LOS MIEMBROS DE
JURADO DE EXAMEN**

Por medio del presente manifiesto que he revisado el texto que será sometido a su consideración del(a) alumno (a) de Maestría en Ciencias Bioquímicas **RAÚL ANTONIO ORTIZ MERINO**, titulado:

"El interactoma en la red metabólica de *Saccharomyces cerevisiae*. Análisis de centralidad y evaluación funcional de las enzimas indispensables para el crecimiento"

ATENTAMENTE



DR. GABRIEL DEL RÍO GUERRA
Tutor Académico



AGRADECIMIENTOS

Esta tesis se realizó bajo la supervisión del Dr. Gabriel del Río Guerra y el Comité Tutor integrado por la Dra. María Alicia González Manjarrez, el Dr. Luis Antonio Mendoza Sierra y el Dr. Alexander de Luna Fors (como miembro invitado) con apoyo por parte del Programa de Maestría y Doctorado en Ciencias Bioquímicas (PMDCBQ) de la Universidad Nacional Autónoma de México (UNAM), del Consejo Nacional de Ciencia y Tecnología (CONACyT) y del Programa de Apoyo a Proyectos de Investigación e Innovación Tecnológica (PAPIIT; proyecto numero IN205911).

El trabajo experimental fue llevado a cabo en el laboratorio de "Integrative Research on Biological Systems" (IRonBioS; oriente 205 IFC-UNAM) con la asistencia técnica de la Dra. María Teresa Lara Ortiz además del personal de la Unidad de Biología Molecular (IFC-UNAM) y de la Unidad de Síntesis y secuenciación de ADN del Instituto de Biotecnología (IBT-UNAM, campus Cuernavaca). Asimismo, se extiende el agradecimiento a las Técnicas Académicas Cristina Aranda, Laura Kawasaki, Olivia Sánchez y Claudia Rodríguez (de los grupos de la Dra. Alicia González, del Dr. Roberto Coría, del Dr. Jesús Aguirre y de Dimitris Georgellis PhD, respectivamente) por su asistencia técnica además del equipo y materiales facilitados.

El trabajo computacional se benefició de las labores de mantenimiento realizadas por la Unidad de Cómputo (IFC-UNAM) a la que también se extiende un agradecimiento.

RESUMEN

La caracterización funcional del genoma de *Saccharomyces cerevisiae* reveló que ~20% de sus genes son indispensables para el crecimiento en medio rico y algunos estudios han tratado de explicar tal fenotipo en relación a la centralidad o importancia relativa de estos genes dentro de la estructura de distintas redes biológicas. Por ejemplo, al analizar redes que relacionan genes a través de las proteínas que producen y que utilizan metabolitos en común, no fue posible encontrar algún índice de centralidad cuyos valores permitieran distinguir los genes indispensables para el crecimiento (**GICs**). Sin embargo, en este trabajo se propone que tales relaciones entre genes que producen proteínas involucradas en reacciones enzimáticas no son suficientes para representar el fenotipo de los GICs. Es por ello que aquí se construyeron nuevas redes de interacciones diversas sobre el metabolismo (**RIDMs**) buscando completar las redes de interacciones metabólicas (**RIMs**) al combinarlas con varias redes de interacciones diversas (**RIDs**) generando así nuevas estructuras que permitan clasificar GICs usando uno o más índices de centralidad.

Este procedimiento resultó en 1311 pares RIDM-centralidad (82.76% de 1584 totales) con los que se pudieron clasificar GICs de manera estadísticamente confiable (usando intervalos de confianza de 99% para el área bajo la curva ROC). Lo anterior en contraste con los resultados previamente reportados por nuestro grupo en donde solo se encontraron 14 casos de 198 (7.07% del total) usando las mismas RIMs. Además, se encontró que en el 14% de estos pares RIDM-centralidad la mejoría en la clasificación de GICs fue debida al incremento en el valor de centralidad por la unión de RIMs y RIDs sin disminuir la capacidad que ya se tenía en la RIM original. Este efecto se presentó tras la integración de cualquiera de las RIMs con alguna de las RIDs que presentan interacciones proteína-proteína (**IPP**) analizadas usando centralidades locales que sólo toman en cuenta las conexiones de los GICs a uno y dos nodos de distancia permitiendo su clasificación. Así, se logró caracterizar la indispensabilidad de hasta 109 GICs de 134 totales (tomando el mejor de los pares RIDM-centralidad) al construir redes conectando genes que producen proteínas relacionadas al metabolismo dependiendo de su participación en vías metabólicas en común según la base de datos KEGG además de las IPP obtenidas en un experimento a gran escala y representadas en la red aquí llamada Yipd. Entre estos 109 GICs existen 56 que solo se pueden clasificar como indispensables al incluir las IPP que establecen sus productos proteicos además de sus relaciones mediadas por metabolitos sugiriendo que estas últimas no eran suficientes. Esto resalta la existencia de proteínas con por lo menos dos funciones distintas (*i.e.* multifuncionales) que son indispensables para el crecimiento y que co-existen durante un mismo proceso celular como es el metabolismo. En este sentido, se encontró que la probable multifuncionalidad en 31 de estos 56 GICs multifuncionales clasificados por aumento en centralidad no puede ser explicada por la simple presencia de más de un dominio funcional según la base de datos PFAM. Por ello, se revisó el papel de algunas de estas enzimas indispensables para el crecimiento en cuanto a los resultados de nuestro análisis y la literatura disponible para analizar más a detalle algunas de estas predicciones. Además, se eligió un GIC de entre los casos discutidos buscando una aproximación experimental útil para analizar la función biológica de este tipo de genes que pudiera ser utilizada posteriormente para validar estas conclusiones.

Así, se mejoró la clasificación de GICs que participan en el metabolismo integrando redes con distintos tipos de interacciones. Los resultados obtenidos con estas nuevas redes concuerdan con la relación entre centralidad y letalidad reportada originalmente en redes de IPP tanto de levadura como de otros organismos. Además, muestran por primera vez que se pueden identificar genes/proteínas indispensables para el crecimiento a partir de la estructura de redes que consideran sus múltiples interacciones dentro de múltiples niveles de organización relacionados con el metabolismo. Así, estas redes incluyen nociones de multifuncionalidad y modularidad que se discuten en relación al estudio de la función biológica y su evolución como propiedades emergentes de los seres vivos al ser considerados como sistemas complejos.

Contenido

I.	INTRODUCCIÓN	1
II.	MATERIALES Y MÉTODOS	4
	Estrategias <i>in silico</i>	4
	II.a REDES Y BASES DE DATOS.....	4
	II.b CENTRALIDADES	6
	II.c DESEMPEÑO DE CENTRALIDADES EN LA CLASIFICACIÓN DE GICs.....	7
	II.d ÍNDICE CIC DE CLASIFICACIÓN POR INCREMENTO DE CENTRALIDAD.....	8
	II.e ANÁLISIS DE DOMINIOS FUNCIONALES	9
	II.f MINERÍA DE TEXTOS	10
	Estrategias <i>in vivo</i>	10
	II.g MUTAGÉNESIS SITIO-DIRIGIDA	11
	II.h SELECCIÓN POR ESPORULACIÓN	13
	II.i ELIMINACIÓN POR RECOMBINACIÓN DE FRAGMENTOS OBTENIDOS POR PCR.....	14
III.	RESULTADOS	16
	III.a CONSIDERAR DISTINTOS INTERACTOMAS MEJORA SIGNIFICATIVAMENTE LA CLASIFICACIÓN DE GICs IMPLICADOS EN EL METABOLISMO.....	16
	III.b LA ADICIÓN DE IPP FACILITA LA CLASIFICACIÓN DE GICs IMPLICADOS EN EL METABOLISMO AL AUMENTAR SU CENTRALIDAD LOCAL	18
	III.c ALGUNOS GICs PRODUCEN PROTEÍNAS MULTIFUNCIONALES.....	20
	III.d HACIA EL DESARROLLO DE UNA ESTRATEGIA <i>IN VIVO</i> PARA EXPLICAR LA INDISPENSABILIDAD DE LOS GICs MULTIFUNCIONALES.....	24
IV.	DISCUSIÓN	29
	IV.a ESTRUCTURA Y FUNCIÓN: REDES, CENTRALIDADES E INDISPENSABILIDAD	29
	IV.b IPP ENTRE ENZIMAS.....	32
	IV.c ENZIMAS MULTIFUNCIONALES INDISPENSABLES PARA EL CRECIMIENTO	34
	IV.d ANÁLISIS <i>IN VIVO</i>	40
	IV.e OTRAS LIMITACIONES Y ALTERNATIVAS.....	43
V.	CONSIDERACIONES FINALES	46
VI.	CONCLUSIONES	48

I. INTRODUCCIÓN

Saccharomyces cerevisiae es un hongo unicelular utilizado en la producción alimentaria (e.g., de pan, cerveza y vino) que además ha sido empleado en la biología molecular y celular desde sus inicios por lo que se encuentra en constante revisión y se ha convertido en un recurso valioso para análisis fisiológicos detallados, particularmente sobre su metabolismo. Esta levadura es un organismo de vida libre con crecimiento rápido que puede mantenerse en grandes poblaciones usando medios de crecimiento relativamente baratos con métodos simples de cultivo que permiten hacer múltiples réplicas para aislar y caracterizar mutantes afectadas en distintos procesos. Asimismo, como se puede encontrar tanto en forma diploide como haploide, permite evaluar fenotipos mediante cruza sexuales y/o división clonal. El uso de este organismo como modelo experimental también ha impulsado el desarrollo de técnicas de manipulación genética que facilitan la expresión de genes ya sea de manera episomal o mediante integración cromosómica simplificando la inserción, eliminación o mutación de casi cualquier secuencia de ADN (Castrillo y Oliver, 2011). Lo anterior permitió eliminar de manera individual la mayoría de las regiones codificantes en el genoma de *S. cerevisiae* para caracterizar efecto en el crecimiento revelando que ~20% de ellas incluyen genes indispensables para el crecimiento (**GICs**) en medio rico (Yeast Peptone Dextrose o **YPD**) (Winzeler, et al., 1999). Se ha planteado que estos genes también llamados esenciales pueden ser utilizados como punto de referencia para entender, modificar y/o controlar procesos celulares por lo que su caracterización constituye una tarea de gran importancia (Chalker y Lunsford, 2002).

Los enfoques de genómica funcional como el que se describe anteriormente han contribuido a que *S. cerevisiae* sea uno de los modelos de estudio preferidos en biología de sistemas junto con otros organismos. Tal es el caso de *Schizosaccharomyces pombe* para el que se cuenta con datos de esencialidad a escala genómica (Kim, et al., 2010). De manera similar, el análisis sistemático de mutaciones de pérdida de función ha conducido a la identificación de casi 3000 genes indispensables para el desarrollo del ratón *Mus musculus* ya sea durante la embriogénesis o después del destete (Skarnes, et al., 2011) y se ha mostrado que varios de ellos se asocian con enfermedades en humanos (Georgi, et al., 2013). Estos estudios que asocian grandes conjuntos de genes con su efecto en la proliferación celular en diferentes condiciones de cultivo han arrojado resultados interesantes que sugieren, por ejemplo, que la alta fracción de genes no indispensables podría reflejar robustez ante mutaciones (Wagner, 2000) o que simplemente no se encuentran activos durante la condición experimental considerada (Blank, et al., 2005). De manera complementaria, se ha reportado que se pueden encontrar vías alternas para sobrellevar la ausencia de proteínas dispensables mientras que las que son indispensables corresponden con la falta de vías alternativas en redes metabólicas perturbadas (Aittokallio y Schwikowski, 2006). Además, la indispensabilidad también ha sido evaluada mediante similitud funcional basada en clasificación ontológica (Ashburner, et al., 2000), en interacciones reportadas en bases de datos (Pagel, et al., 2005) o en la similitud de niveles de coexpresión y perfiles filogenéticos (Zhang, et al., 2012).

La reconstrucción y el modelado *in silico* se han utilizado como un marco consistente dentro del cual formular las relaciones entre GICs y buscar elementos que puedan explicar su función sobre todo en relación con el metabolismo al asociar genes que producen proteínas involucradas en reacciones bioquímicas (Palsson, 2009). Dichas relaciones toman en cuenta los sustratos y productos compartidos por distintas proteínas por lo que también se llaman gen-proteína-reacción. En este sentido, la información genómica, bioquímica y fisiológica sobre el metabolismo de *S. cerevisiae* ya ha sido utilizada para reconstruir distintas representaciones metabólicas (Duarte, et al., 2004; Förster, et al., 2003; Österlund, et al., 2013) y evaluarlas con respecto a datos de crecimiento en cepas mutantes obtenidos por métodos experimentales directos o bien por métodos computacionales (Förster, et al., 2003; Duarte, et al., 2004; Acencio y Lemke, 2009; del Rio, et al., 2009). Sin embargo, en estos trabajos no se ha podido explicar la naturaleza de los GICs aun cuando se enfocan exclusivamente en aquellos genes que participan en el metabolismo de la levadura cuando crece en YPD.

Además, aún se desconoce si el fenotipo que presentan los GICs que participan en el metabolismo depende de la actividad catalítica de la proteína que codifica o de otros tipos de interacciones, como las IPP en las redes de interacciones físicas (Ito, et al., 2001; Uetz, et al., 2000) u otros tipos de relaciones moleculares indirectas como las interacciones genéticas detectadas mediante técnicas como los arreglos sintéticos de genes (del inglés "synthetic genetic arrays" o SGA (Tong, et al., 2001; Ooi, et al., 2003)) o bien mediante los perfiles epistáticos de mini-arreglos (del inglés "epistatic mini-array profiles" o EMAPs (Schuldiner, et al., 2005)). En este sentido, la primer descripción curada manualmente de las interacciones proteicas y genéticas de la levadura *S. cerevisiae* (Reguly, et al., 2006) inició una serie de esfuerzos para convertir interacciones reportadas en la literatura a un formato computable y puso en evidencia que aun en esta levadura el conjunto completo de interacciones (*i.e.* interactoma) aún se desconoce (Dolinski, et al., 2013).

Es por ello que ya se han realizado distintos experimentos computacionales buscando modelar la función de redes de escala celular integrando varias aproximaciones de genómica funcional (Vidal, 2005) y se han comparado con respecto a las que consideran un solo tipo de aproximación (Kohlstedt, et al., 2010; de Matos Simoes, et al., 2013). Tales enfoques se basan en la hipótesis de que los organismos vivos pueden caracterizarse mediante una jerarquía de sistemas con distintos niveles interdependientes (Weiss, 1969) cuyas interacciones podrían revelar principios generales de organización en sistemas biológicos (Mesarović y Takahara, 1972). Así, para tomar en cuenta los distintos niveles de organización biológica que han sido propuestos (Southern, et al., 2008) que ocurren en distintas escalas espaciales y temporales (Walker y Southgate, 2009) es necesario integrar modelos que puedan ser reducidos y/o ampliados (Dada y Mendes, 2011). Dichos modelos han sido llamados multinivel (Mesarović y Takahara, 1972) o multiescala (Southern, et al., 2008) como parte de una perspectiva que ha sido referida como multi-ómica (Liu, et al., 2013). Así, independientemente de cómo se les nombre, los análisis que buscan integrar conjuntos de datos contrastan con aquellos en los que los procesos son analizados utilizando un solo nivel genómico (*e.g.* transcriptoma) oponiéndose al reduccionismo que considera al organismo como una máquina y a sus funciones como mecanismos concentrándose en procesos aislados por lo que pasa por alto elementos biológicos importantes (Woese, 2004; Rosslenbroich, 2011).

Así, el estudio de funciones biológicas (como el fenotipo relacionado con los GICs) podría encontrar utilidad en métodos que permitan acoplar y/o analizar modelos celulares a partir de elementos biológicos asociados mediante diversos tipos de relaciones funcionales a distintas escalas temporales y espaciales. Sin embargo, al asumir tal modelo de organización celular surge la pregunta de cómo es que se pueden determinar los diferentes niveles de subordinación. En primera instancia, no queda claro si estos niveles deberían o no, presentar elementos en distintas escalas temporales o espaciales. Además, una vez definidos los niveles aún sería necesario determinar la relación entre ellos y como es que se coordinan. Es por ello que aún no hay conclusiones definitivas al respecto pero la intuición general es que hacia los niveles de menor orden se podrían describir mejor las relaciones causales a costa de perder contexto mientras que hacia los niveles de mayor orden se podría observar mejor el contexto dificultando definir factores con precisión (Dolinski, et al., 2013). De este modo, una representación completa de la célula eucarionte (y/o de componentes importantes como los GICs dentro de un proceso celular como el metabolismo) debería incluir sus principales niveles de regulación con distintos mecanismos y redes (Castrillo y Oliver, 2011). En este sentido, también se ha postulado que la importancia implícita tras esta búsqueda de niveles y sus relaciones podría revelar principios de organización (Mesarović y Takahara, 1972) involucrando conceptos de evolucionabilidad, propiedades emergentes, adaptación y robustez con sus distintas interpretaciones (Wellstead, et al., 2008). Así, independientemente de cómo se les llame, el fondo teórico detrás de estas aproximaciones integradas recae en la definición de propiedades emergentes en la teoría de sistemas (von Bertalanffy, 1976) que indican cómo es que distintas características de un sistema solo se pueden observar cuando el sistema se estudia como un todo y no como la simple suma de sus partes, acepción cada vez más popular en el ámbito de la biología.

Desde tal punto de vista, en este trabajo se utilizaron interacciones 'químico-genéticas' para representar la relación entre los genes que se expresan como enzimas y utilizan metabolitos en común ya sea como sustratos o como productos. También se consideraron otros tipos de relaciones funcionales

directas, como las interacciones físicas del tipo proteína-proteína, o indirectas, como las obtenidas por dobles eliminaciones y epistásis, e incluso algunas inferencias computacionales. Así, se construyeron redes en donde se asociaron distintos genes con base en las proteínas que estos producen y en las relaciones que estas establecen de manera que el tipo de interacción es un reflejo de su función. De tal forma, las interacciones de los genes más centrales y la medida de centralidad con la que fueron clasificados como indispensables pueden ser utilizados para plantear hipótesis buscando explicar su fenotipo. Con ello se mejoró la clasificación de GICs que están implicados en el metabolismo, con respecto a esfuerzos anteriores, y se encontró que algunos GICs codifican proteínas cuya actividad enzimática es suficiente para explicar su fenotipo mientras que algunos otros GICs sólo se pueden clasificar como indispensables al considerar relaciones que van más allá de su función enzimática (e.g. interacciones proteína-proteína). Este último conjunto sugiere la existencia de proteínas con más de una función (*i.e.* multifuncionales) y su descripción podría ser útil para estudiar ciertos atributos propios a los sistemas biológicos que no son predecibles a partir de interacciones o componentes individuales que son conocidas como propiedades emergentes. Por ello también se revisó el papel de algunas de estas enzimas indispensables para el crecimiento en cuanto a los resultados de nuestro análisis y la literatura disponible para analizar mas a detalle algunas de estas predicciones. Además, se eligió un GIC de entre los casos discutidos buscando una aproximación experimental útil para analizar la función biológica de estos GICs que producen proteínas multifuncionales y validar las conclusiones que aquí se presentan.

II. MATERIALES Y MÉTODOS

El estado incompleto del conocimiento biológico actual se puede notar de distintas maneras, por ejemplo, aún se desconoce si la indispensabilidad de ciertos genes depende de las interacciones que establecen con otros genes (o sus productos proteicos) indispensables o no. Tales relaciones incluyen las interacciones proteína-proteína (**IPP**) o las interacciones proteína-ADN además de las llamadas químico-genéticas (o relaciones gen-proteína-reacción) en las que se asocian genes que se expresan como enzimas que utilizan un mismo metabolito como sustrato o producto. Así, este trabajo pretende relacionar el fenotipo de los genes indispensables para el crecimiento (**GICs**) de *S. cerevisiae* con su importancia dentro de la estructura de distintas redes construidas a partir de diversos tipos de interacciones entre diferentes moléculas. Hasta donde se sabe, aún no se ha intentado probar distintas combinaciones de representaciones metabólicas, interactomas y medidas de centralidad para optimizar la clasificación de los GICs que participan en el metabolismo de *S. cerevisiae* como se describe a continuación.

Los métodos y los materiales aquí descritos se basan en la hipótesis de que la indispensabilidad de los genes que participan en el metabolismo no radica solamente en la función catalítica de las enzimas para las que codifican. Por tal motivo, se utilizaron enfoques *in silico* e *in vivo* de manera complementaria y se describen con detalle en las secciones correspondientes.

Estrategias *in silico*

II.a REDES Y BASES DE DATOS

De acuerdo con el paradigma estructura-función, un fenotipo puede ser representado como una red R con n genes definida por un conjunto de genes V (vértices o nodos) y un conjunto de relaciones A (aristas). De tal manera se pueden definir los conjuntos:

$$\begin{aligned} R &= (V, A) \\ V(R) &= \{v_1, v_2, \dots, v_n\} \\ A(R) &= \{v_1v_2, v_2v_4, \dots, v_xv_y\} \end{aligned}$$

que no pueden ser conjuntos vacíos y en donde n es el número total de genes de la red. Dentro de tales conjuntos se pueden definir los GICs como el conjunto de nodos o vértices $I(R)$ de manera que $I(R) \in V(R)$; es decir, $V(R)$ incluye aquellos genes indispensables que son identificables al ser ordenados usando los valores arrojados por distintas operaciones matemáticas (*i.e.* medidas de centralidad).

Estas definiciones permitieron la elaboración de programas implementando operaciones para construir redes y los algoritmos para analizarlas. Así se generaron 18 Redes de Interacciones Metabólicas (**RIMs**) según el método reportado anteriormente (del Rio, et al., 2009), 8 de ellas a partir de la base de datos KEGG (Ogata, et al., 1999) y otras 10 que son variantes de la red iND750 reconstruida usando datos experimentales (Duarte, et al., 2004). Estas RIMs fueron obtenidas conectando los genes a través de los metabolitos que comparten (*e.g.* cuando un gen codifica una enzima que utiliza un compuesto químico que fue producido por otra enzima codificada por otro gen). Después, se generaron distintas versiones de cada red (ver **Tabla 1**) en donde se eliminaron los metabolitos más conectados y/o los genes hipotéticos (utilizados al representar reacciones que aún no han sido asociadas con genes o enzimas particulares) además de tomar en cuenta la reversibilidad de la reacción, su participación en vías previamente descritas y compartimentación.

Adicionalmente, los genes de *S. cerevisiae* cuya interacción se reportó en 7 bases de datos distintas se utilizaron como vértices para elaborar redes donde la existencia de interacción da lugar a las aristas (**Tabla 2**). Con esto se generaron 8 redes de interacciones diversas (**RIDs**) llamadas así debido a que incluyen relaciones entre genes que producen proteínas que interactúan físicamente (interacciones proteína-proteína) o genes a los que les fue asociada una relación funcional indirecta (interacciones genéticas). Tales interacciones fueron obtenidas tanto mediante técnicas de alto rendimiento (SGA y dobles híbridos a escala genómica) como por experimentos individuales e incluso algunas de ellas fueron asociadas a partir de su aparición en el mismo texto (co-citación en la literatura) o curadas manualmente como se indica en la **Tabla 2** según sea el caso. Adicionalmente, se construyó una octava red que proviene del conjunto unión de las interacciones reportadas en las 7 bases de datos referidas en la **Tabla 2** y se utilizó para evaluar el efecto de todas las interacciones diversas sobre las distintas RIMs independientemente de en cuantas bases de datos fueran reportadas. También se intentó construir el conjunto intersección para evaluar solamente aquellas interacciones reportadas en todas las bases de datos pero este conjunto resultó vacío.

Tabla 1

Red	Metabolitos eliminados	Vías	EC	Genes hipotéticos removidos	Compartimientos
KEGG	H ₂ O, ATP, ADP, NAD ⁺ , NADH, NADP, NADPH	si	si	ND	ND
KEGGtype	H ₂ O, ATP, ADP, NAD ⁺ , NADH, NADP, NADPH	si	si	ND	ND
KEGGpath	H ₂ O, ATP, ADP, NAD ⁺ , NADH, NADP, NADPH	si	si	ND	ND
KEGGtypepath	H ₂ O, ATP, ADP, NAD ⁺ , NADH, NADP, NADPH	si	si	ND	ND
KEGG2	H ₂ O, ATP, ADP, NAD ⁺ , NADH, NADP, NADPH	si	ND	ND	ND
KEGG2type	H ₂ O, ATP, ADP, NAD ⁺ , NADH, NADP, NADPH	si	ND	ND	ND
KEGG2path	H ₂ O, ATP, ADP, NAD ⁺ , NADH, NADP, NADPH	si	ND	ND	ND
KEGG2typepath	H ₂ O, ATP, ADP, NAD ⁺ , NADH, NADP, NADPH	si	ND	ND	ND
iND750_H_0	H ₂ O, H ⁺	si	ND	no	Si
iND750_H_1	H ₂ O, H ⁺ , Pi	si	ND	no	Si
iND750_H_2	H ₂ O, H ⁺ , Pi, ATP	si	ND	no	Si
iND750_H_3	H ₂ O, H ⁺ , Pi, ATP, Glu-L	si	ND	no	Si
iND750_H_4	H ₂ O, H ⁺ , Pi, ATP, Glu-L, ADP	si	ND	no	Si
iND750_nH_0	H ₂ O, H ⁺	si	ND	si	Si
iND750_nH_1	H ₂ O, H ⁺ , Pi	si	ND	si	Si
iND750_nH_2	H ₂ O, H ⁺ , Pi, ATP	si	ND	si	Si
iND750_nH_3	H ₂ O, H ⁺ , Pi, ATP, Glu-L	si	ND	si	Si
iND750_nH_4	H ₂ O, H ⁺ , Pi, ATP, Glu-L, ADP	si	ND	si	Si

RIMs. EC: Número EC de clasificación enzimática. ND: no determinado. ADP: Difosfato de Adenosina; ATP: Trifosfato de Adenosina; Glu-L: L-Glutamato; H⁺: Protón; H₂O: agua; NAD⁺: Nicotinamida Dinucleótido Oxidada; NADH: Nicotinamida Dinucleótido Reducida; NADP: Nicotinamida Dinucleótido-Fosfato; NADPH: Nicotinamida Dinucleótido-Fosfato; Pi: Fosfato.

Tanto las RIMs como las RIDs fueron generadas como listas de adyacencia (representando un elemento por columna cuya adyacencia en la misma fila representa una interacción) en archivos de texto delimitados por tabuladores que contienen dos columnas. Estas listas se elaboraron en a partir de la información descargada de las bases de datos de la **Tabla 2** tomando solo las interacciones donde ambos identificadores pertenecen a *S. cerevisiae*. Estas redes de interacciones funcionales no tienen dirección de manera que si el gen A interactúa con el gen B también puede decirse lo contrario, es decir, que el gen B interactúa con el gen A. Esto es debido a que en los distintos métodos utilizados para construir las bases de datos de la **Tabla 2** no se especifica la direccionalidad de la interacción; por tanto, todas las redes utilizadas a continuación incluyen el total de interacciones reportadas en la base de datos correspondiente y son consideradas en ambas direcciones independientemente de en qué dirección fueron descritas. Por ejemplo, si la interacción A-B y la interacción B-A se encuentran en la base de datos son incluidas en la red al igual que las relaciones de B-C y C-B aun cuando C-B no haya sido reportada. Además, si las aristas de cada red se definen según el par de nodos u, v , si $u=v$ se trata de un "self-loop", si no, se define como una arista propia. Estos "self-loops" fueron retirados de cada red debido a que varios de los algoritmos con los que se obtienen los valores de centralidad requieren del cálculo de las distancias más cortas mediante el algoritmo de Dijkstra que no converge si la red incluye "self-loops".

Tabla 2

Base de datos	IG	IPP	AR	CCL	CM	Sitio web
Biogrid	X	X	X	X	X	http://www.thebiogrid.org/
Dip		X		X		http://dip.doe-mbi.ucla.edu/
Intact		X		X	X	ftp://ftp.ebi.ac.uk/pub/databases/intact/current
Mpact		X		X	X	ftp://ftp.mips.gsf.de/fungi/yeast/PPI/
Yeastnet	X					http://www.yeastnet.org/
Yipd		X	X			http://itolab.cb.k.u-tokyo.ac.jp/Y2H/
Ypi		X	X			http://depts.washington.edu/sfields/yp_interactions/index.html

RIDs. AR: alto rendimiento; CCL: co-citación en la literatura; CM: curada manualmente; IG: interacciones genéticas; IPP: interacciones proteína-proteína

Se construyeron 144 redes de interacciones diversas en el metabolismo (**RIDMs**) que representan el conjunto unión para cada una de las 18 diferentes RIMs (KEGG, KEGGpath, KEGGtype, KEGGtypepath 1 y 2 además de iND750_H e iND750_NH de la _0 a la _4) con las 8 RIDs. La transformación de las bases de datos en listas y en redes además de las operaciones de conjuntos entre ellas se realizaron mediante programas escritos en Perl. Todas estas redes junto con sus respectivos componentes gigantes (CGs; refiriéndose al subconjunto de vértices conectados mediante por lo menos una arista) se analizaron mediante distintos criterios de centralidad y se evaluaron en cuanto a su utilidad para clasificar genes indispensables como se describe en las secciones siguientes.

II.b CENTRALIDADES

Las medidas de centralidad están definidas dentro de la teoría de redes como algoritmos que evalúan la importancia relativa de un nodo dentro de la estructura de una red. En este caso, los genes de cada red se ordenaron de mayor a menor valor según 10 medidas de centralidad (**Tabla 3**) y se relacionaron con listas de GICs para evaluar el desempeño predictivo de cada centralidad sobre cada red (ver sección **II.c**). Además de los valores de centralidad para cada gen se calcularon el diámetro (distancia máxima entre cualquier par de nodos), el parámetro OCCI (Ma y Zeng, 2003) que representa la distribución de los valores de la cercanía (referida en la tabla como “Clos”) y el promedio de las centralidades CC, Clos, Ecc, <Dist>, Deg y SD (identificadas como meanClusteringCoef, meanClosenessCent, meanExcentricity, meanDistance, meanDegree y meanSphereDegree respectivamente).

Adicionalmente, se obtuvo un índice combinando los valores de 2, 3 y 4 medidas de centralidad para cada gen, nodo o vértice usando la siguiente fórmula:

$$IC(v) = \left(\sum_{i=1}^m (\text{MAX}_{C_i} - C(v))^2 / (\text{MAX}_{C_i} - \text{MIN}_{C_i}) \right)^{1/2}$$

donde MAX_{C_i} y MIN_{C_i} definen, respectivamente, el valor máximo y mínimo obtenido para la centralidad i , $C(v)$ se refiere a la centralidad de un gen dado dentro de una red y el índice m se refiere al orden de las combinaciones de centralidades de manera que $m=\{1,2\}$ para grupos de 2 centralidades, $m=\{1,2,3\}$ para grupos de 3 centralidades y $m=\{1..4\}$ para grupos de 4 centralidades. Así, $IC(v)$ es un índice que incluye los valores de las centralidades para el gen analizado y estima que tan lejos están de los valores de centralidad más altos, de tal manera, mientras más bajo sea el valor de $IC(v)$ mayor es el valor individual de las centralidades consideradas. Con esto se obtuvieron 45 grupos de 2 centralidades, 120 grupos de 3 y 210 de 4, los cuáles fueron utilizadas al igual que las 10 centralidades individuales para comparar su uso sobre los distintos tipos de redes (RIMs, RIDs y RIDMs).

Tabla 3

Índice de centralidad	Abreviatura	Descripción
"Clustering coefficient"	CC	Proporción entre el número de conexiones de un nodo y sus vecinos con respecto al número total de conexiones posibles entre vecinos
"Clustering coefficient inverse"	CCinv	Valor inverso aditivo de CC
"Degree"	Deg	Número de interacciones por nodo
"Second Nearest Neighbor"	SNN	Número de interacciones para cada nodo a una distancia igual a 2 de un nodo particular
"Sphere Degree"	SD	Número de interacciones para cada nodo a una distancia menor o igual a 2 de un nodo particular
"Closeness"	Clos	Recíproco de la distancia promedio entre cada nodo y todos los demás en la red
"Eccentricity"	Ecc	Distancia más corta entre un nodo y el nodo más lejano
"Eccentricity inverse"	Eccinv	Valor inverso aditivo de Ecc
"Mean Distance"	<Dist>	El promedio de todas las distancias de un nodo con respecto a todos los demás
"Traversity"	Trav	Conectividad dinámica o el número de veces que un nodo es atravesado para llegar cada par de nodos a través de su vía más corta según su orden dentro de la red

Medidas de centralidad

El cálculo de los valores de centralidad se realizó por medio de programas codificados en Java implementando métodos reportados anteriormente (Cusack, et al., 2007; Thibert, et al., 2005). Las combinaciones de centralidades fueron calculadas según el método reportado anteriormente por nuestro grupo (del Rio, et al., 2009). Los procesos de cálculo tanto de los valores de centralidad como de los IC(v)s fueron automatizados y paralelizados utilizando el modulo Parallel::ForkManager disponible en el CPAN (<http://www.cpan.org/>) de PERL. Los valores de las medidas generales para cada red son presentados como archivo de texto en la Tabla C1 complementaria.

II.c DESEMPEÑO DE CENTRALIDADES EN LA CLASIFICACIÓN DE GICs

Los genes considerados como GICs son los que aparecen como esenciales en el inventario reportado en el sitio http://www-sequence.stanford.edu/group/yeast_deletion_project/deletions3.html por el *Saccharomyces* Genome Deletion Project y fueron utilizados para obtener listas de genes indispensables a partir de cada una de las redes. En este estudio se considera que los GICs contenidos en las RIMs están involucrados en el metabolismo y se expresan como enzimas indispensables para el crecimiento. Para mantener esta relación, en el caso de las RIDMs solo se consideraron los GICs contenidos en la RIM correspondiente. Lo anterior no es posible con las RIDs para las cuales se utilizaron los GICs presentes en la red aún cuando no necesariamente están relacionados al metabolismo.

Para evaluar el desempeño clasificatorio de las distintas medidas de centralidad sobre cada red se utilizaron varios parámetros que dependen del número de verdaderos positivos (**VP**), falsos positivos (**FP**), verdaderos negativos (**VN**) y falsos negativos (**FN**). Estos grupos son definidos en términos de los fenotipos medidos experimentalmente y del grupo en el que fueron clasificados. Así, los VP son los genes reportados como indispensables que fueron clasificados como GICs mientras que los FP son los genes cuya eliminación produce células viables pero fueron clasificados como GICs. Los negativos se definen de manera complementaria por lo que los VN son los genes cuya eliminación produce células viables por lo que no fueron clasificados como GICs y los FN son los GICs que no se encuentran en el grupo clasificado como indispensable.

Dichos parámetros permiten calcular la sensibilidad ($VP/(VP+FN)$), la especificidad ($VN/(VN+FP)$), la exactitud ($(VP+VN)/(VP+VN+FP+FN)$) y la tasa de falsos positivos ($FP/(VP+FP)$) que son útiles en la construcción de curvas receptor-operador (**ROC** por sus siglas del inglés "Receiver Operating Characteristic") que se utilizan comúnmente para evaluar el desempeño de métodos predictivos o de clasificación. En este estudio se calculó el área bajo la curva ROC (**ABC**) para reducir cada curva a un solo índice cuantitativo y facilitar así su comparación. Tal índice varía desde 0.5 (sin exactitud aparente)

hasta 1.0 (perfectamente exacto) y fue utilizado como medida de la probabilidad con la que se distinguen los GICs de los genes que no son indispensables para el crecimiento. Para cuantificar la variabilidad en los valores de ABC se utilizó el error estándar calculado a partir de dos probabilidades intermedias Q1 (probabilidad de que dos mediciones negativas tomadas al azar sean clasificadas mejor que una positiva escogida aleatoriamente) y Q2 (probabilidad de que una medición negativa escogida al azar sea mejor clasificada que dos mediciones positivas escogidas aleatoriamente) para distinguir las probabilidades α y β de cometer errores tipo I o tipo II, respectivamente, y construir así intervalos de confianza. Así, el ABC acompañada por su error estándar permitió la construcción de intervalos de confianza (IC) de 0.1, 0.05 y 0.01 (90%, 95% y 99%) como pruebas para obtener la significancia estadística. Lo anterior implica que un modelo efectivo deberá generar un valor de ABC significativamente mayor que 0.50 (valor esperado al azar) de manera equivalente a la estadística de Wilcoxon o a la U de Mann Whitney normalizada por el número de pares posibles (Hanley y McNeil, 1982). El valor de error mínimo para el ABC distingue el valor de centralidad con el que se obtuvieron los valores máximos de especificidad y sensibilidad de manera conjunta lo que sirvió para rastrear el punto de la curva más cercano a la predicción perfecta (Wunderlich y Mirny, 2006).

Este análisis fue realizado mediante programas codificados en Java implementando el método desarrollado en nuestro grupo. De igual manera que con el cálculo de centralidades (ver sección **CENTRALIDADES**), se utilizó el módulo Parallel::ForkManager disponible en el CPAN (<http://www.cpan.org/>) para acelerar el cálculo de ABC. La estadística, tanto descriptiva como inferencial para los datos obtenidos, fue realizada mediante funciones básicas del software estadístico R (Becker, et al., 1988) y de algunos paquetes adicionales disponibles en el CRAN (<http://cran.r-project.org/>) indicados a continuación. Para representar los datos se elaboraron gráficas de tipo "boxplot" usando la función con el mismo nombre (*boxplot*). Los datos se describieron mediante funciones del paquete *psych*. Las pruebas de normalidad Shapiro-Wilk y Anderson-Darling así como la prueba Fligner-Killeen para la homogeneidad de las varianzas y la estadística no-paramétrica de Wilcoxon para las diferencias entre los dos grupos provienen del paquete *nortest*. Las tablas utilizadas como "input" se elaboraron mediante "scripts" de AWK y/o Perl cuyo "output" se redirigió tanto a figuras como a archivos de texto. Todos los valores de ABC para cada RIM usando cada medida de centralidad de manera individual o combinada son presentados como archivo de texto en la Tabla C2 complementaria. La Tabla C3 complementaria incluye los mismos valores en el mismo formato que la Tabla C2 pero para las RIDMs.

II.d ÍNDICE CIC DE CLASIFICACIÓN POR INCREMENTO DE CENTRALIDAD

Para distinguir los GICs que incrementaron su valor de centralidad tras la adición de interacciones aumentando con ello la probabilidad de clasificarlos como tales es necesario que cumplan lo siguiente:

- 1) que existan GICs predichos como indispensables solo en la RIDM y no en la RIM original, y
- 2) que tal clasificación se deba a un aumento en su valor de centralidad en la RIDM con respecto a la RIM original.

Para ello, se obtuvieron todos los VP con valor de centralidad mayor o igual a aquel con el que se obtuvo el valor mínimo de error con cada par RIM-centralidad y cada par RIDM-centralidad por separado. De tal modo, el conjunto de GICs que fueron clasificados correctamente en una RIDM (por ejemplo Yipd.KEGG2) sin considerar el conjunto de GICs predichos correctamente usando la misma centralidad en la RIM correspondiente (KEGG2 siguiendo el ejemplo) son aquellos que cumplen con el inciso 1) y son denotados como **GICs'**. Para obtener los GICs' que además cumplen con el inciso 2), se calculó la diferencia DC del valor de centralidad obtenido en la RIM original (CRIM) menos el valor obtenido en la RIDM (CRIDM) correspondiente con la misma centralidad para cada uno de estos genes de manera que:

$$DC_{ij} = CRIM_{ij} - CRIDM_{ij} \begin{cases} < 0; \text{ el valor de centralidad aumentó al añadir interacciones} \\ = 0; \text{ el valor de centralidad no cambió al añadir interacciones} \\ > 0; \text{ el valor de centralidad disminuyó al añadir interacciones} \end{cases}$$

Donde $CRIM_{ij}$ y $CRIDM_{ij}$ corresponden a los valores de centralidad obtenidos con un índice de centralidad i para cada GIC' j . Así, aquellos GICs' cuya DC es negativa implican que el valor de centralidad obtenido usando la RIDM en cuestión aumentó con respecto al obtenido usando la RIM correspondiente.

Con estos valores se construyó un índice de clasificación por incremento en centralidad (**CIC**) para cada red en el que el número de GICs' que cumplen con los incisos 1) y 2) son evaluados con respecto al número total de GICs' como sigue:

$$CIC = \frac{\{GICs' | DC < 0\}}{\{GICs' | DC < 0\} + \{GICs | DC = 0\} + \{GICs | DC > 0\}}$$

o simplemente:

$$CIC = \frac{\{GICs' | DC < 0\}}{\{GICs\}}$$

revelando la proporción de GICs' correctamente clasificados solo en la RIDM y no en la RIM correspondiente cuya centralidad aumentó con respecto al total de GICs. Así, mientras el CIC sea mayor a 0 y aumente hacia valores cercanos a 1, mayor será la proporción de GICs' predichos por un incremento en su valor de centralidad debido a la adición de nuevas interacciones predichos con la RIDM y la centralidad en cuestión. Entonces, las redes con $ABC > 0.5$ y $CIC > 0$ son aquellas con las que los GICs se pueden clasificar correctamente y de manera confiable debido al aumento en el valor de centralidad seguido por la adición de nuevas interacciones. Todos los valores de ABC, CIC y otros parámetros útiles para cada una de las combinaciones red-centralidad analizadas son presentados como archivo de texto en la tablaC4 complementaria.

II.e ANÁLISIS DE DOMINIOS FUNCIONALES

Los 134 GICs descritos por el "Saccharomyces Genome Deletion Project" (http://www-sequence.stanford.edu/group/yeast_deletion_project/deletions3.html) contenidos en al menos una de las RIMs se analizaron en cuanto a su contenido de dominios funcionales. Para ello se obtuvo la secuencia de aminoácidos en formato FASTA para cada proteína codificada por cada GIC en cuestión utilizando el nombre del ORF correspondiente en la herramienta YeastMine (<http://yeastmine.yeastgenome.org/yeastmine/>) del "Saccharomyces genome database" (**SGD**; (Cherry, et al., 2012)). Utilizando dichas secuencias de aminoácidos se realizó la búsqueda de dominios en cualquiera de los servidores de PFAM (<http://pfam.sanger.ac.uk/>, <http://pfam.sbc.su.se/> y <http://pfam.janelia.org/>; (Punta, et al., 2012)). Dicha búsqueda fue ejecutada por un programa escrito en JAVA para acceder a dichos servidores y obtener los dominios PFAM para cada secuencia de aminoácidos una por una de manera que fue automatizado para analizar las 134 secuencias y para arrojar el output en formato HTML mediante programas escritos en Perl. Así, una vez obtenidos los dominios PFAM para cada uno de los GICs, éstos fueron contrastados con la lista obtenida en el sitio: ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/complete/docs/yeast.txt que contiene el nombre de cada gen con todos sus sinónimos y los nombres de los ORFs utilizados por el

proyecto de secuenciación del genoma en cuestión, el nombre ordenado de cada locus, el número primario de acceso en UniProtKB/Swiss-Prot (<http://www.ebi.ac.uk/uniprot>), el nombre de entrada, el número de acceso en la base de datos del genoma de *S. cerevisiae* (<http://www.yeastgenome.org/>), la longitud de la secuencia, la disponibilidad de estructura tridimensional y la localización cromosómica para el proteoma completo de *S. cerevisiae*.

II.f MINERÍA DE TEXTOS

Se utilizó el motor de búsqueda de PubMed (<http://www.ncbi.nlm.nih.gov/pubmed>) para encontrar los artículos que hacen referencia a enzimas y al concepto "moonlighting" usando el término de búsqueda moonlight* AND ((protein* OR enzym*) OR (activit* OR function*)). Los PMIDS (identificadores de la base de datos PubMed) y los artículos correspondientes obtenidos con el término de búsqueda fueron curados a mano para detectar proteínas multifuncionales y los genes que las codifican. Adicionalmente, se obtuvo el texto en formato XML de los títulos y resúmenes correspondientes a estos PMIDS. Tal información fue analizada mediante el API (del inglés "Application Programming Interface" o interfaz de programación de aplicaciones) para JAVA del software LingPipe (<http://alias-i.com/lingpipe/index.html>) en donde, entre otras cosas, se pueden separar objetos del tipo proteína anotados por el proyecto GENIA (<http://www.nactem.ac.uk/genia/>). Dichos genes, proteínas e identificadores fueron contrastados mediante programas escritos en PERL con la lista del proteoma de *S. cerevisiae* descrita en la sección anterior, con la lista de GICs del SGDP y con las listas de GICs correctamente clasificados para saber si pertenecen a esta levadura, si son indispensables para el crecimiento y si fueron correctamente clasificados entre las proteínas identificadas como multifuncionales.

Estrategias *in vivo*

En este caso se eligió el GIC codificado por el **ORF** (del inglés "Open Reading Frame" o marco abierto de lectura) *YDR050C* para diseñar una estrategia con la cual analizar la función biológica de este tipo de genes y validar el método aquí presentado. Este gen fue elegido debido a que se encontró entre los GICs asociados a más de una función (**Tabla 8**) obtenidos mediante los métodos descritos en las secciones **II.a** a **II.e** del apartado de **Estrategias *in silico***. Además, este GIC produce una de las proteínas cuyo estudio ha sido central para el desarrollo de la enzimología mecanicista (Rieder y Rose, 1959) por lo que su sitio activo ya ha sido descrito en varias ocasiones (Komives, et al., 1991; Go, et al., 2010). De manera importante, también se ha discutido que la falta de su actividad enzimática no es letal en humanos (Ralser, et al., 2006; Orosz, et al., 2009) y en *Drosophila melanogaster* se demostró *in vivo* que su deficiencia puede ser complementada con una mutante sin actividad catalítica (Roland, et al., 2013) apoyando la hipótesis de este trabajo en donde la función enzimática parece no ser suficiente para explicar su indispensabilidad. Este gen también presenta un fenotipo de letalidad sintética cuando se elimina junto con el ORF *YHR129C* (Biogrid), que codifica para la proteína ARP1 relacionada con la actina, lo que apoya aún más su multifuncionalidad en la levadura. Además, se reportó que la sustitución del ácido glutámico número 97 por un residuo de glutamina (E97Q) abate la actividad de la Triosa Fosfato Isomerasa (TPI) de *Plasmodium falciparum* sin presentar modificaciones estructurales globales adicionales (Samanta, et al., 2011). Por tanto, se compararon las secuencias de aminoácidos de las proteínas con actividad de TPI1 en *P. falciparum* (XP_001348552.1) y de *S. cerevisiae* (NP_010335.1) usando la versión 2.2.29 el programa BLASTP (Altschul, et al., 1997; Altschul, et al., 2005) con el que se pueden alinear con una cobertura del 99%, identidad del 42% y un E-value de 2×10^{-75} (lo suficientemente pequeño con respecto al valor máximo de 10 para indicar un buen alineamiento). Además, se compararon 2 de las estructuras cristalográficas reportadas en el banco de datos de proteínas (o PDB del inglés Protein Data Bank) para las proteínas correspondientes de ambos organismos (1NEY *S. cerevisiae* para y 1O5X para *P. falciparum*) utilizando la versión 3.1 del programa DaliLite para comparaciones en

pares (Hasegawa y Holm, 2009) y se encontró identidad del 41% entre ambas con Z-scores de hasta 35.5 (donde cualquier alineamiento con Z-score menor a 2 es considerado espurio). En adelante se muestran las estrategias utilizadas para, a partir del GIC *YDR050C*, generar una proteína con la mutación E97Q sin actividad catalítica manteniendo su estructura general para evaluar el efecto de la actividad catalítica en relación a su indispensabilidad para el crecimiento de *S. cerevisiae* (ver secciones **IV.c** y **IV.d** para una discusión más a detalle).

II.g MUTAGÉNESIS SITIO-DIRIGIDA

La mayoría de los ORFs de *S. cerevisiae* están disponibles en una colección de plásmidos llamada MoBY ORFs incluyendo desde ~900 pares de bases río arriba del codón de inicio hasta ~250 pares de bases río abajo del codón de paro de cada gen en la secuencia genómica de la cepa S288C (Ho, et al., 2009). Dichos vectores de expresión fueron construidos a partir del plásmido p5472 por lo que pueden ser utilizado para expresar genes en levadura. Además, al ser introducido en la cepa BUN20 de *Escherichia coli* permite mantener y recuperar estos plásmidos a partir de células cultivadas en medio LB bajo en sales (**Tabla 6**) para, subsecuentemente, purificarlo por métodos rápidos tales como el paquete "miniprep" de QIAGEN. Así, fue obtenido el plásmido que es llamado MoBY_TPI1_wt en adelante con el ORF *YDR050C* correspondiente al GIC responsable de la producción de la proteína TPI1 con actividad de Triosa-fosfato isomerasa (Open Biosystems, número de catálogo YSC3867-9522828, localización 40-F-5; Donación del Dr. Alexander de Luna; ver **Tabla 4**). Una vez recuperado se digirió con la enzima *Eco*NI (New England) para obtener un fragmento lineal cuyo peso molecular fue verificado por electroforesis y comparado con los patrones de restricción obtenidos con la enzima *Xba*I (Invitrogen). Los sitios de corte para ambas enzimas se muestran en el mapa de la **Figura 1**. También se utilizaron los oligos para confirmación descritos en la **Tabla 5** diseñados para amplificar por PCR una región de ~400pbs a partir de la región promotora del gen *YDR050C* para confirmar la identidad del plásmido MoBY_TPI1_E97Q de una manera alternativa. Tal reacción se llevó a cabo utilizando 30 ciclos considerando 45 segundos a 60°C para alinear los oligos y 3 minutos de elongación a 68°C usando la polimerasa *Taq* (New England) suplementada con MgCl₂ (50mM).

Posteriormente, el plásmido MoBY_TPI1_wt se utilizó para generar la mutación E97Q descrita por Moumita Samanta y colaboradores (Samanta, et al., 2011). Dicha mutación fue realizada utilizando el paquete "QuickChange lightning" (Stratagene) de mutagénesis sitio-dirigida con modificaciones en el paso de transformación que se describen más adelante. Siguiendo las especificaciones de tal paquete, se diseñó un par de oligonucleótidos sintéticos de 41 bases que contienen únicamente la mutación deseada justo a la mitad y flanqueada por 10 bases de secuencia complementaria al gen de TPI1 correcta a ambos lados, presentando T_m=80.60, contenido de GC=43.90%, mismatch de 2.44 y al menos 3 residuos C o G en los extremos (ver secuencia en **Tabla 5**). Estos oligonucleótidos mutagénicos fueron diseñados para provocar la mutación G1137C (según su posición en el plásmido) cambiando el codón GAA que originalmente codifica un residuo de ácido glutámico por el codón CAA que codifica para glutamina y es usado preferencialmente en *S. cerevisiae* (18.5 veces por cada mil vs 1.4 veces por cada mil con el codón CAG). Estos oligonucleótidos fueron sintetizados por la unidad de Biología Molecular en el Instituto de Fisiología Celular, UNAM y sus características se muestran en la **Tabla 5**. La posición de la mutación insertada se muestra en el mapa del plásmido MoBY_TPI1_wt de la **Figura 1**.

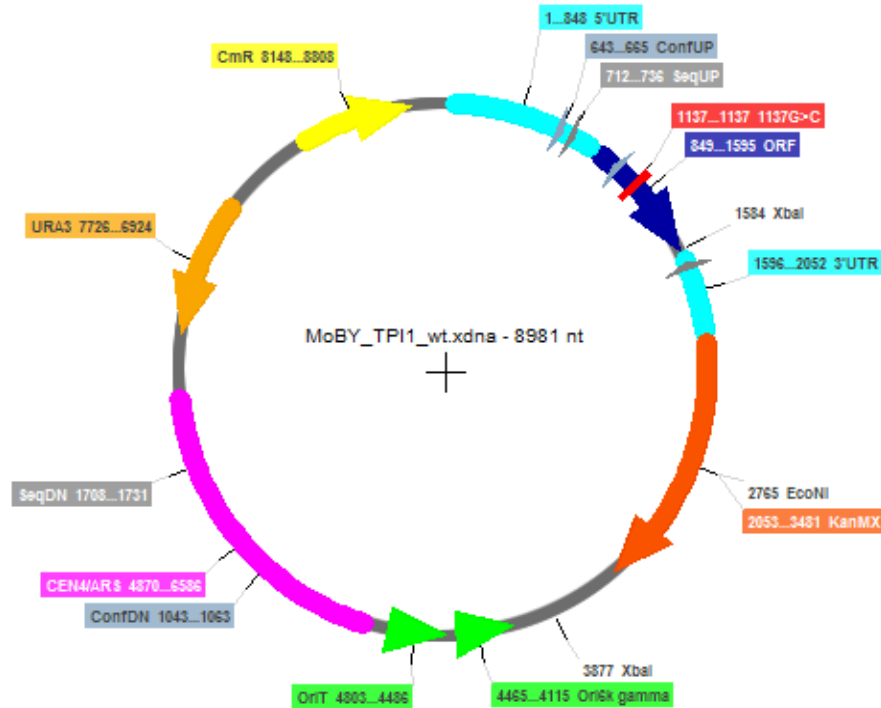


Figura 1. Mapa del plásmido MoBY_TPI1. Se indica la posición del ORF correspondiente al gen *YDR050C* responsable de la producción de la proteína TPI1p con actividad de Triosa Fosfato Isomerasa. Este ORF está acompañado de sus regiones 5' y 3' no codificantes, además se indica el sitio en donde se produjo la mutación E97Q y la región de homología con los oligonucleótidos de secuenciación. También se muestran los marcadores de selección CmR que confiere resistencia a Cloranfenicol para su selección en bacteria además de KanMX y URA3 para su selección en levadura. En morado se muestra el sitio CEN4/ARS de reconocimiento por el centrosoma de levadura. Los sitios marcados como OriT y Ori6k gamma denotan la posición del origen de transferencia y el origen de replicación respectivamente.

Es importante resaltar que los plásmidos de la colección MoBY fueron construidos con el método MAGIC que utiliza los procesos de conjugación y recombinación bacteriana para evitar el uso de enzimas de restricción y/o recombinasas sitio-específicas (Li y Elledge, 2005). Como consecuencia dichos plásmidos presentan el sitio *OriT* y otros elementos del sistema de transferencia del factor F bacteriano y solo pueden replicarse en células que presentan el gen *pir1* (en este caso el alelo *pir1-116*) que codifica el factor π que actúa en trans sobre el origen de replicación *ori6k γ* (Metcalf, et al., 1994). Ante la falta de acceso a la cepa BUN20 (**Tabla 4**) utilizada originalmente en la colección MoBY, el material genético obtenido con la reacción de mutagénesis descrita en el párrafo anterior se verificó por electroforesis y se utilizó para transformar la cepa BW25142 (donación del Dr. Dimitris Georgellis; ver genotipo en **Tabla 4**) que cuenta con el genotipo *pir-116* a diferencia de la cepa XL10 gold distribuida con el "kit" de mutagénesis. Adicionalmente, la cepa BW25142 se transformó por electroporación, no por choque térmico como es indicado en tal "kit". La selección del plásmido con la mutación G1137C fue realizada en medio LB (**Tabla 6**) suplementado con kanamicina (100 μ g/ml), y cloranfenicol (12.5 μ g/ml). Dos de éstas colonias se recuperaron para purificar el plásmido por "miniprep" y verificar la mutación determinando la secuencia de nucleótidos del gen *YDR050C* en dicho vector. Tal secuencia se obtuvo en la Unidad de Biología Molecular del Instituto de Fisiología Celular a partir de un fragmento que abarca 137 pbs río arriba del gen y hasta 136 pbs río abajo. Para ello se utilizaron los oligonucleótidos de secuenciación UP y DN (**Tabla 5**) que fueron sintetizados en la Unidad de Síntesis y secuenciación de ADN del Instituto de Biotecnología (UNAM, campus Cuernavaca). El plásmido MoBY con el gen de la TPI1 que presenta la mutación E97Q se refiere en adelante como MoBY_TPI1_E97Q.

II.h SELECCIÓN POR ESPORULACIÓN

Las cepas Y26690 y Y23986 de la **Tabla 4** provienen de la cepa diploide BY4743 de *S. cerevisiae* con el genotipo MAT a/ α ; *YDR050C::kanMX4/YDR050C* donde la secuencia del ORF correspondiente se intercambi3 por un cassette KanMX4 que confiere resistencia a geneticina mediante la misma estrategia de remoci3n basada en PCR y recombinaci3n utilizada por el SGDP (Giaever, et al., 2002; Winzeler, et al., 1999). Estas c3lulas diploides heterocigotas fueron sometidas a varios procedimientos buscando obtener mutantes nulas para el gen *YDR050C*. Las dos estrategias iniciales para tal prop3sito incluyeron transformaci3n, esporulaci3n y selecci3n de esporas con distintas particularidades y en distinto orden en secuencia (**Figura 2**).

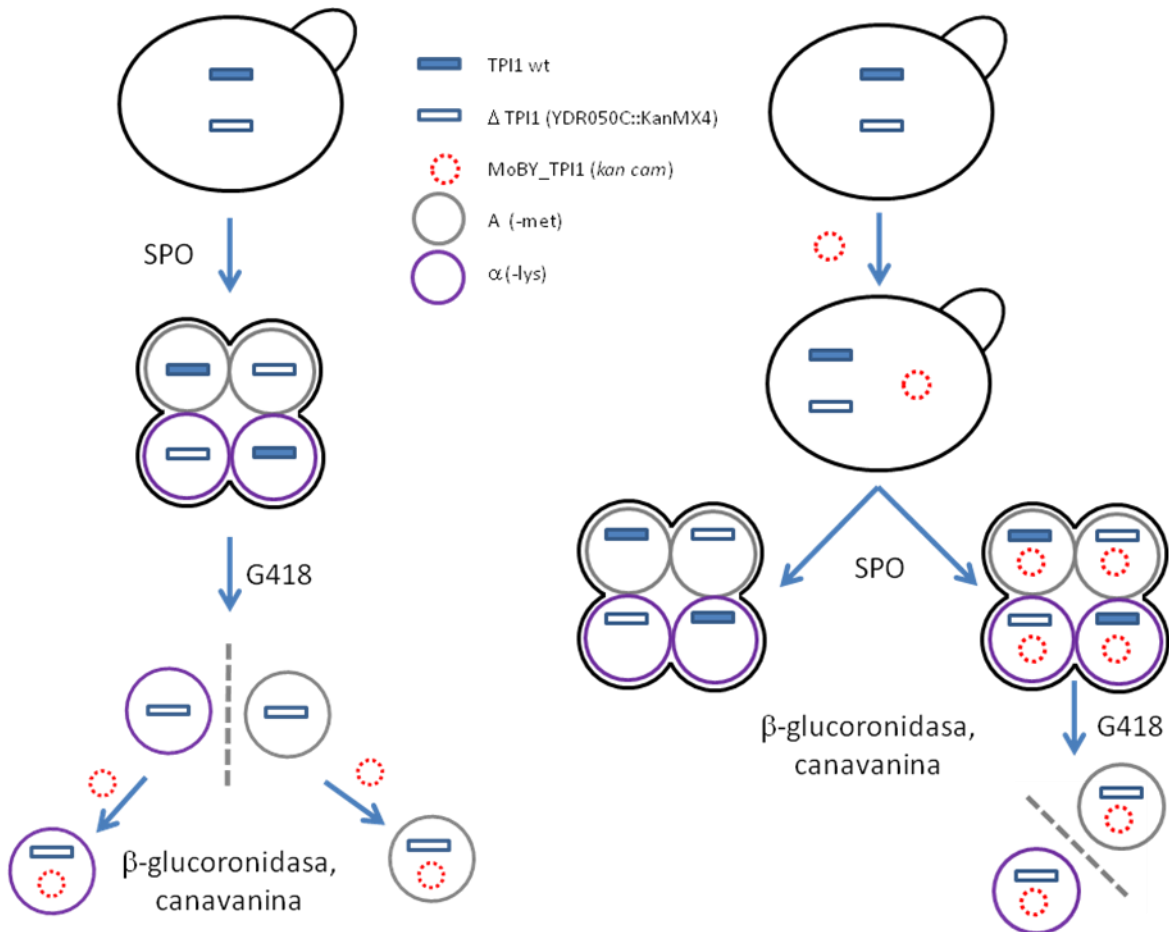


Figura 2 Estrategias alternativas para seleccionar células haploides nulas para el gen *YDR050C* con el plásmido MoBY silvestre o mutante. La estrategia de la izquierda parte de las cepas diploides heterocigotas que son sometidas a esporulaci3n para seleccionar las células nulas y despu3s transformarlas con el plásmido MoBY en cuesti3n. En la derecha tambi3n se inicia con las células diploides heterocigotas pero éstas son transformadas en primera instancia con el plásmido MoBY en cuesti3n y despu3s se seleccionan las células haploides que sean nulas para el gen *YDR050C* y tambi3n contengan el plásmido MoBY en cuesti3n. El c3digo de colores y formas se muestra en la parte central.

La primera estrategia involucr3 la transformaci3n de las células diploides heterocigotas con los plásmidos MoBY_TPI1_wt o MoBY_TPI1_E97Q seguida por la selecci3n de esporas nulas para el gen *YDR050C* con el vector en cuesti3n. La segunda estrategia implica la selecci3n de esporas nulas para el gen *YDR050C* seguida por la transformaci3n con el plásmido MoBY_TPI1_wt o MoBY_TPI1_E97Q de las células haploides nulas recuperadas. As3, tanto el plásmido MoBY_TPI1_wt como aquel con la

mutación E97Q se utilizaron para transformar las células provenientes de las cepas Y26690 y Y23986 por separado siguiendo el protocolo de Chen modificado por Burke (Chen, et al., 1992; Burke, et al., 2005). Estas cepas transformadas con los distintos plásmidos se sometieron a un procedimiento experimental modificado (Giaever, et al., 2002) para hacerlas esporular con el fin de segregar las mutaciones de interés (Burke, et al., 2005). Después de confirmar la obtención de ascosporas mediante microscopía óptica, los cultivos fueron concentrados por centrifugación y se les añadieron perlas de vidrio y 500 unidades de β -glucuronidasa (tipo H3; Sigma) para después someterlas a una incubación en rotor por una hora a 30 °C. Esta suspensión de células diploides y ascosporas digeridas fue sembrada en medio SC -arg adicionado con canavanina (60 μ g/ml; Sigma) para seleccionar las células haploides en dos lotes ya sea -lys o -met. Además, los medios de selección fueron adicionados con G418 (300 μ g/ml; Gibco) para seleccionar el genotipo *YDR050C::kanMX4* y con inositol (75-100 μ M; Sigma) para complementar el crecimiento de las cepas sin TPI1 acorde a lo reportado para las cepas SMY10 y SMY15 (Shi, et al., 2005). Es importante mencionar que en cualquiera de los casos en los que se utilizó medio SC con G418 el glutamato monosódico al 1% fue la fuente de nitrógeno debido a que el sulfato de amonio impide la acción de antibióticos como el G418 y el clonNAT (Yan Tong y Boone, 2005). Los componentes de estos medios de cultivo se describen en la **Tabla 6**. Las colonias obtenidas recuperadas tras la selección por esporulación fueron analizadas por PCR a partir de colonias (Burke, et al., 2005) utilizando los oligonucleótidos de secuenciación UP y DN (**Tabla 5**).

II.i ELIMINACIÓN POR RECOMBINACIÓN DE FRAGMENTOS OBTENIDOS POR PCR

Se utilizó el protocolo de Gietz (Gietz y Woods, 2002) semejante al utilizado por el SGDP (Giaever, et al., 2002; Winzeler, et al., 1999) en el que se utilizan fragmentos de PCR obtenidos de tal manera que presentan homología con los extremos de algún gen de interés pero en su lugar contienen un "cassette" que confiere resistencia al ser utilizados para transformar cultivos de células que integran dicho fragmento eliminando el gen silvestre. Para ello se emplearon los oligos de secuenciación UP y DN descritos en la **Tabla 5** además de los fragmentos generados con los oligos del gen de STE2 como controles positivos. Ya que la ausencia del gen *YDR050C* impide el crecimiento en medio rico, las células transformantes fueron recuperadas en medio con inositol como se indica en la sección anterior para complementar el crecimiento de las cepas sin TPI1 acorde a lo reportado para las cepas SMY10 y SMY15 (Shi, et al., 2005), en medio YPEG para la cepa W303 (Compagno, et al., 1996) y en medio YPED para la cepa CEN.PK 113-7D (CBS8340) (Compagno, et al., 2001). La composición de los distintos medios de cultivo se muestra en la **Tabla 6**.

II.j CEPAS, OLIGONUCLEÓTIDOS Y MEDIOS DE CULTIVO

Tabla 4

Organismo	ID	Genotipo
<i>E. coli</i>	BUN20	Δ lac-169 rpoS(Am) robA1 creC510 hsdR514 Δ uidA(MluI)::pir-116 endA(BT333) recA1 F'(lac ⁺ pro ⁺ Δ oriT::tet)
	BW25142	F', Δ (araD-araB)567, Δ lacZ4787(::rrnB-3), Δ (phoB-phoR)580, λ , galU95, Δ uidA4::pir-116, recA1, endA9(del-ins)::FRT, rph-1, Δ (rhaD-rhaB)568, hsdR514
<i>S. cerevisiae</i>	BY4741	MAT α ; his3 Δ 1; leu2 Δ 0; met15 Δ 0; ura3 Δ 0
	BY4742	MAT α ; his3 Δ 1; leu2 Δ 0; lys2 Δ 0; ura3 Δ 0
	BY4743	MAT α /MAT α his3 Δ 0/his3 Δ 0; leu2 Δ /leu2 Δ 0; met15 Δ 0/MET15; LYS2/lys2 Δ 0; ura3 Δ 0/ura3 Δ 0
	Y26690	MAT α /MAT α Δ <i>YDR050C::kanMX4/ YDR050C</i>
	Y23986	MAT α /MAT α Δ <i>YDR050C::kanMX4/ YDR050C</i>
	Δ STE2	BY4741 Δ <i>YFL026W::kanMX4</i>

Cepas. Se muestra el genotipo de cada cepa, ya sea de *E. coli* o *S. cerevisiae*, utilizada en este estudio.

Tabla 5

Tipo		Secuencia 5'-3'
Mutagénicos	UP	GGGTTATTTTGGGTCCTCCaaAGAAGATCTTACTTCCAC
	DN	GTGGAAGTAAGATCTTCTTtgGGAGTGACCCAAAATAACCC
Secuenciación	UP	AACTTGCAACATTTACTATTTTCCC
	DN	AGAACATTACGAAATTTAAGTGCC
Confirmación TPI	UP	CCTATATACCTTTGGCTCGGCTG
	DN	CCAGAAGCCTTCAAGTAGGCG
STE2	UP	GATACCTTTTCTTTTCACCTGC
	DN	ATGTGGTGCATCTGATGAGC

Oligonucleótidos. Las secuencias de nucleótidos mostradas en esta tabla fueron sintetizadas y utilizadas como templados en las distintas reacciones de PCR descritas tanto en los métodos como en los resultados.

Tabla 6

Medio	Descripción (cantidades por litro de agua destilada)
Luria-Bertani	Extracto de levadura 5g, Bacto-triptona 10g, NaCl 5g
YPD	Extracto de levadura 1%; Peptona 1%; Dextrosa 2%
GNA de preesporulación	Extracto de levadura 0.8%; Peptona 0.3%; Dextrosa 5%
Esporulación	KOAc 1%; Dextrosa 0.025%
YPED	Extracto de levadura 1%; Peptona 2%; Etanol 2%; Dextrosa 0.1%
YPEG	Extracto de levadura 1%; Peptona 2%; Etanol 3%; Galactosa 0.1%
SC-ura	Dextrosa 2%; Yeast Nitrogen Base 6.7g; fosfato de potasio monobásico (KH ₂ PO ₄) 10g; sulfato de amonio ((NH ₄) ₂ SO ₄) 50g; Drop-out mix -ura 0.77 g
SC-lys	Dextrosa 2%; Yeast Nitrogen Base 6.7 g; fosfato de potasio monobásico (KH ₂ PO ₄) 10g; sulfato de amonio ((NH ₄) ₂ SO ₄) 50g; Drop-out mix -lys 0.77 g
SC-met	Dextrosa 2%; Yeast Nitrogen Base 6.7 g; fosfato de potasio monobásico (KH ₂ PO ₄) 10g; sulfato de amonio ((NH ₄) ₂ SO ₄) 50g; Drop-out mix -met 0.77 g
SC-arg	Dextrosa 2%; Yeast Nitrogen Base 6.7 g; fosfato de potasio monobásico (KH ₂ PO ₄) 10g; sulfato de amonio ((NH ₄) ₂ SO ₄) 50g; Adenina 0.001%; Histidina 0.002%; Leucina 0.010%; Triptófano 0.005%; Uracilo 0.002%; Metionina 0.002%; Lisina 0.005%
SC+ino	*SC-ino ; inositol 10mM
*SC-ino	**Vitaminas 1ml; ***Elementos traza 1ml; ****Sales y fuente de nitrógeno 100ml; Dextrosa 20g; Adenina 0.001%; Arginina 0.005%; Histidina 0.002%; Leucina 0.010%; Triptófano 0.005%; Uracilo 0.002%; Metionina 0.002%; Lisina 0.005%
	**Vitaminas: Biotina 2mg; pantotenato de calcio 400mg; ácido fólico 2mg; niacina 400mg; ácido para-aminobenzoico 200mg; hidrocloreto de piridoxina 400mg; robloflavina 200mg; hidrocloreto de tiamina 400mg (NaOH para disolver)
	***Elementos traza: Ácido bórico (H ₃ BO ₃) 0.5g; sulfato de cobre (CuSO ₄) 0.04g; yoduro de potasio (KI) 0.1g; cloruro férrico (FeCl ₃) 0.2g; sulfato de manganeso (MnSO ₄) 0.4g; molibdato de sodio (NaMoO ₄) 0.2g; sulfato de zinc (ZnSO ₄) 0.4g
	****Sales y fuente de nitrógeno: Fosfato de potasio monobásico (KH ₂ PO ₄) 10g; sulfato de magnesio (MgSO ₄) 5g; cloruro de sodio (NaCl) 1g; cloruro de calcio (CaCl ₂) 1g; glutamato monosódico 20g

Medios de cultivo. La cepas descritas en la **Tabla 4** fueron cultivadas en los medios cuya composición se muestra aquí.

III. RESULTADOS

Se utilizaron redes genéticas del metabolismo reconstruidas mediante relaciones químico-genéticas de acuerdo al procedimiento descrito anteriormente (del Rio, et al., 2009) que a su vez fueron enriquecidas con diversos tipos de interacciones. Lo anterior con la idea de complementar con la información faltante que podría permitir explicar la indispensabilidad de la función de los GICs en el metabolismo de la levadura *S. cerevisiae*. La información contenida en dichas redes fue obtenida tanto de la base de datos KEGG (Ogata, et al., 1999) como del modelo iND750 (Duarte, et al., 2004) (ver **Tabla 1**; sección **II.a**), en tanto que las interacciones moleculares se obtuvieron de bases de datos públicas (**Tabla 2**; sección **II.a**). Posteriormente, se utilizó el conocimiento obtenido al distinguir la(s) centralidad(es) más útil(es) en la identificación de genes indispensables para plantear hipótesis sobre el papel que juegan estos genes dentro de la estructura de las distintas redes y a partir de ello tratar de explicar la causa de la indispensabilidad de estos genes tanto por métodos *in silico* como *in vivo*.

III.a CONSIDERAR DISTINTOS INTERACTOMAS MEJORA SIGNIFICATIVAMENTE LA CLASIFICACIÓN DE GICs IMPLICADOS EN EL METABOLISMO

Se había reportado que al combinar centralidades se puede mejorar la identificación de GICs a partir de sus valores de centralidad (del Rio, et al., 2009). Acorde a lo anterior, las mejores clasificaciones obtenidas con las RIMs (Redes de Interacciones Metabólicas) fueron resultado de la combinación de medidas de centralidad según el índice descrito en la sección **II.b**. En el panel A de la **Figura 3** se puede notar que las clasificaciones obtenidas al combinar 2 o más centralidades usando el método descrito en la sección **II.c** sobre las RIMs son mejores que al utilizar centralidades individuales. Específicamente, el mejor de los casos resultó de combinar 3 distintas medidas de centralidad; sin embargo, solo permitió identificar el 61.4% de los GICs (iND750_NH_4 SD-clos-trav ABC=0. 65±0.048, ver valores en Tabla C2 complementaria).

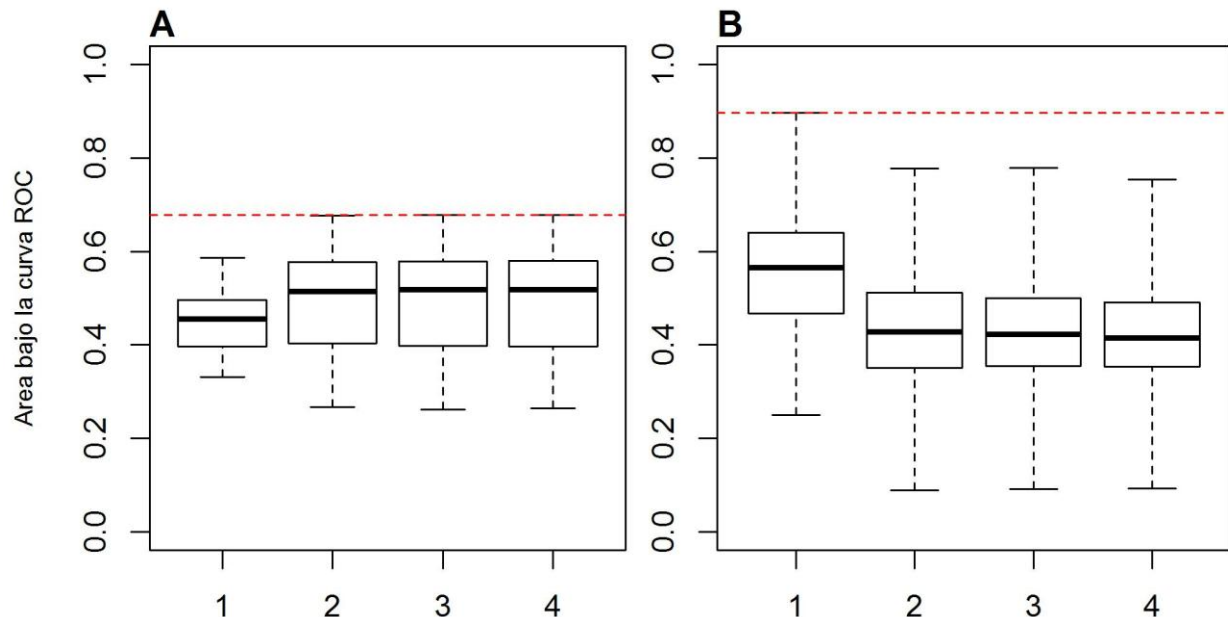


Figura 3. Se utilizaron 10 diferentes índices de centralidad combinados en grupos de 2, 3 y 4 centralidades (indicados de manera respectiva en el eje de las ordenadas) sobre 18 RIMs (panel A) y 144 RIDMs (panel B) para evaluar su utilidad en la clasificación de GICs según el área bajo la curva ROC (ABC; eje de las ordenadas). Las centralidades también fueron. Los boxplots resumen la distribución de valores de ABC para cada par red-centralidad representando la mediana con la línea gruesa central alrededor de la cual se dibuja la caja que incluye el segundo y tercer cuantiles mientras que las líneas punteadas hacia los extremos de cada caja se extienden hacia los valores máximo y mínimo. La línea roja discontinua marca el valor máximo de ABC.

Las interacciones entre los genes de esta levadura que fueron reportadas en 7 diferentes bases de datos (**Tabla 2**; sección **II.a**) se utilizaron para construir nuevas redes, una a partir de cada base de datos más su conjunto unión que son referidas como RIDs (Redes de Interacciones Diversas). Estas 8 RIDs fueron unidas con las 18 RIMs construidas previamente lo que resultó en 144 redes nuevas referidas como RIDMs. Tales reconstrucciones fueron sometidas al mismo método para calcular centralidades tanto individuales como combinadas y les fue aplicado el mismo análisis para determinar las mejores combinaciones entre redes y centralidades en la identificación de GICs.

Así, en el panel B de la **Figura 3** se puede notar que la integración de los diversos tipos de interacciones presentes en las RIDs sobre las RIMs resultó en valores más altos de ABC que aquellos obtenidos con las RIMs originales usando tanto centralidades individuales como de manera combinada (ver valores en la Tabla C3 complementaria). Esto muestra que la integración de nuevas interacciones en las mencionadas RIDMs resulta más útil en la clasificación de GICs que considerar centralidades de manera combinada. Además, los valores de ABC obtenidos con las RIDMs no presentan la misma mejoría que en el caso de las RIMs al considerar las centralidades en forma combinada. De esta manera, el supuesto de que la clasificación de GICs a partir de RIMs no es posible debido a que estas redes están incompletas parece más adecuado que aquel en el que se considera que los grupos de centralidades reflejan la naturaleza compleja de las interacciones entre genes del metabolismo.

En la **Figura 4** se presenta una comparación entre las ABC obtenidas al utilizar las diferentes medidas de centralidad para clasificar los genes indispensables de las 18 RIMs, las 8 RIDs y las 144 RIDMs (ver valores en la Tabla C4 complementaria). Además se compararon los resultados obtenidos usando la subred conexas de mayor tamaño o componente gigante (CG), para los tres tipos de redes (evaluadas según los conjuntos de genes críticos correspondientes, ver detalles en la sección **II.c**). Al realizar este procedimiento se podrían esperar hasta 15 pares red-centralidad con valores de $ABC > 0.5$ (según un modelo aleatorio nulo con 99% de confianza sobre un total de 1440 pruebas realizadas sobre las RIDMs); sin embargo, se presentaron 1311 pares red-centralidad con $ABC > 0.5$ en contraste con las 14 de 198 posibilidades (7.07% del total) en las que se presentó un $ABC > 0.5$ usando RIMs.

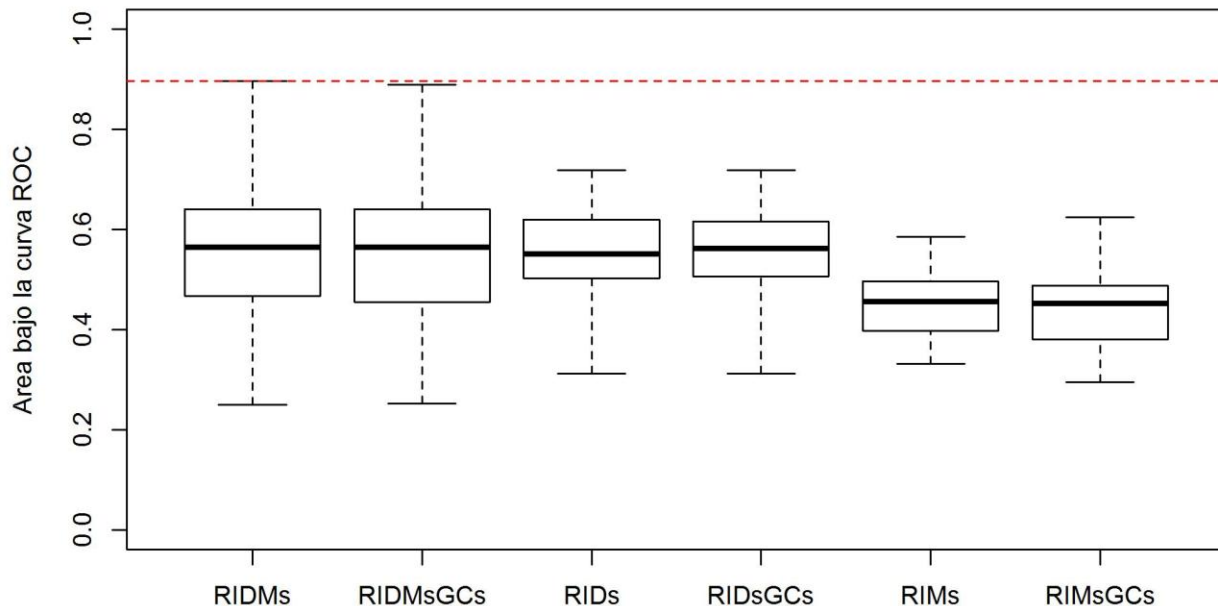


Figura 4. Las mismas 10 medidas de centralidad fueron utilizadas sobre 340 redes (18 RIMs, 8 RIDs, 144 RIDMs y sus 170 componentes gigantes o CGs) para identificar los genes indispensables para el crecimiento (GICs). La probabilidad de discriminar correctamente los genes indispensables de los que no lo son fue evaluada utilizando el área bajo la curva ROC (ABC). Los boxplots resumen el valor de ABC para cada par red-centralidad con la mediana representada en la línea media o segundo cuartil mientras que los bigotes van hacia los valores máximo y mínimo respectivamente. La línea roja discontinua marca el valor máximo de ABC indicando el mejor de todos los resultados obtenidos.

Para comparar los distintos parámetros estadísticos obtenidos según los métodos descritos en las secciones **II.b** y **II.c** tanto para las RIMs como para las RIDMs se analizó a detalle la distribución de los valores de ABC obtenidos con ambos grupos. De esta manera fue posible detectar que las RIDMs son hasta un 30% mejores que las RIMs al obtenerse valores de ABC de hasta 0.9. Además, la distribución de valores de ABC obtenidos con las RIDMs presenta un sesgo negativo, lo que indicó cierta tendencia hacia los valores más altos. Dicha tendencia se confirmó al encontrar una diferencia significativa entre la distribución de ABC obtenidas con las RIMs y las RIDMs ($p < 2.2 \times 10^{-16}$) según la prueba no paramétrica de Wilcoxon. Se utilizó dicha prueba debido a que los valores de ABC, tanto de las RIMs como de las RIDMs, no presentan distribuciones de tipo normal según las pruebas Shapiro-Wilk ($p = 2.574 \times 10^{-8}$ para RIDMs y $p = 1.283 \times 10^{-3}$ para RIMs) y Anderson-Darling ($p < 3.212 \times 10^{-12}$ para RIDMs y $p = 0.009544 \times 10^{-3}$ para RIMs) además de que sus varianzas no son homogéneas según la prueba Fligner-Killeen ($\chi^2 = 74.935$, $p < 2.2 \times 10^{-16}$). Adicionalmente, el uso de los CGs de las RIMs generó algunos resultados mejores que los obtenidos para las respectivas redes completas y no ocurrió lo mismo al evaluar los CGs de las RIDMs. Esto resalta desde una segunda perspectiva la importancia de incluir las interacciones de las RIDs para completar las RIMs lo cual mejoró la clasificación de los GICs. Con esto se corroboró que las RIDMs son más útiles para distinguir los GICs que cualquiera de las RIMs o RIDs originales por separado.

III.b LA ADICIÓN DE IPP FACILITA LA CLASIFICACIÓN DE GICs IMPLICADOS EN EL METABOLISMO AL AUMENTAR SU CENTRALIDAD LOCAL

Hasta este punto los resultados obtenidos apoyan el supuesto de que al considerar distintos tipos de interacciones es posible obtener una representación del metabolismo que permite la clasificación de GICs a partir de medidas estructurales. Para cumplir con lo anterior, la mejoría en la clasificación debería ser consecuencia del aumento en el valor de centralidad de los GICs contenidos en las distintas RIMs tras la adición de las interacciones de las diferentes RIDs en las nuevas RIDMs. Por ello se construyó el índice CIC de clasificación por incremento en centralidad (ver sección II.d) con el que se pudo notar que 735 de los 1440 pares RIDM-centralidad presentaron la condición de $ABC > 0.5$ y $CIC > 0$ presentando un valor máximo de $ABC = 0.755 \pm 0.037$ obtenido por la red Intact.Palsson_HIPOT_4 usando la centralidad Trav (ver tabla C4 complementaria; ver Palsson_HIPOT_4 en **Tabla 1**, Intact en **Tabla 2** y Trav **Tabla 3**). Esta RIDM solo incluye 100 GICs de los cuales 58 fueron clasificados correctamente con la red Palsson_HIPOT_4 aumentando a 74 cuando fueron clasificados usando esta misma centralidad ya en la RIDM. En este caso, 29 GICs fueron clasificados exclusivamente en la RIDM y todos ellos aumentaron su valor con respecto al obtenido usando la centralidad Trav sobre la RIM correspondiente por lo que obtuvieron un $CIC = 1$. Pese a lo anterior, 13 GICs dejaron de ser correctamente clasificados tras la adición de interacciones, es decir, se predijeron en la RIM pero no en la RIDM (**Tabla 7** Sección B).

Tomando en cuenta estas observaciones se identificaron 206 pares RIDM-centralidad con los que además de cumplir la condición de $ABC > 0.5$ y $CIC > 0$ ningún GIC dejó de ser correctamente clasificado al añadir interacciones (como ocurrió en el caso de la red Intact.Palsson_HIPOT_4). Entre este último grupo el valor más alto de ABC fue de 0.753 ± 0.024 usando la centralidad SD sobre la red Yipd.KEGG clasificando 97 GICs de 134 totales entre los cuales 38 fueron predichos exclusivamente en la RIDM y 59 ya eran clasificados con el par RIM-centralidad correspondiente (**Tabla 7** Sección C). Además, la diferencia entre los valores de ABC obtenidos con el par Intact.Palsson_HIPOT_4-Trav (**Tabla 7** Sección B) y con el par Yipd.KEGG-SD (**Tabla 7** Sección C) no es significativa aún cuando con este último par se predicen 23 GICs más con probabilidad semejante y cumpliendo la condición de $ABC > 0.5$ y $CIC > 0$ (presentando valores de $CIC = 1$ en ambos casos). Así, en el 14% de los pares RIDM-centralidad la mejoría en la predicción de GICs fue debida al incremento en el valor de centralidad por la unión de RIMs y RIDs sin perder la capacidad predictiva que ya se tenía en la RIM. Con esto se pudo notar que aunque aparecieron todas las RIMs solo se encontraron las RIDs Intact, Mpact, Yipd, Ypi y Union todas ellas presentando IPP. Además, en este grupo aparecieron todas las centralidades locales a excepción de 12 casos en los que fue utilizada la centralidad Ecclnv y 3 en los que se utilizó Trav (**Tabla 7** Secciones C y D) señalando que las IPP añadidas a estas redes permiten la clasificación de GICs sobre todo mediante centralidades locales.

Tabla 7

etiqueta	ABC	99%ls	99%li	error	sensibilidad	especificidad	CIC	GICs	RIMVPs	RIDMVPs	DC<0	DC>0	DC=0	{RIMVPs- RIDMVPs}
A														
Yipd.KEGG2-CC	0.896	0.920	0.872	0.182	0.896	0.852	0.000	134	71	120	0	24	25	0
Yipd.KEGG-CC	0.895	0.919	0.870	0.181	0.903	0.847	0.000	134	71	121	0	26	24	0
Yipd.KEGG2path-CC	0.892	0.917	0.866	0.186	0.903	0.841	0.000	134	69	121	0	23	29	0
Yipd.KEGGpath-CC	0.890	0.915	0.865	0.186	0.903	0.841	0.000	134	71	121	0	23	27	0
Mpact.KEGG2-CC	0.865	0.891	0.839	0.258	0.843	0.795	0.000	134	71	113	0	28	14	0
Yipd.KEGG2-Deg	0.860	0.882	0.838	0.286	0.791	0.805	0.548	134	64	106	23	0	19	0
Yipd.KEGG-Deg	0.860	0.882	0.838	0.296	0.776	0.807	0.585	134	63	104	24	0	17	0
Mpact.KEGG-CC	0.862	0.888	0.836	0.267	0.851	0.779	0.000	134	71	114	0	30	14	1
Mpact.KEGG2path-CC	0.859	0.886	0.832	0.269	0.821	0.800	0.000	134	69	110	0	28	15	2
Yipd.KEGG2path-Deg	0.852	0.875	0.829	0.304	0.806	0.766	0.558	134	56	108	29	0	23	0
B														
Intact.Palsson_HIPOT_4-Trav	0.756	0.793	0.718	0.401	0.740	0.694	1.000	100	58	74	29	0	0	13
Intact.KEGG-Deg	0.763	0.790	0.735	0.431	0.716	0.675	1.000	134	63	96	36	0	0	3
Intact.KEGG2-Deg	0.762	0.790	0.735	0.433	0.716	0.673	1.000	134	64	96	35	0	0	3
Intact.Palsson_HIPOT_3-Trav	0.751	0.789	0.714	0.408	0.740	0.685	1.000	100	55	74	29	0	0	10
Intact.Palsson_HIPOT_1-Trav	0.752	0.789	0.714	0.406	0.703	0.723	1.000	101	54	71	30	0	0	13
Intact.Palsson_HIPOT_0-Trav	0.752	0.789	0.715	0.408	0.706	0.718	1.000	102	58	72	28	0	0	14
Intact.Palsson_HIPOT_2-Trav	0.751	0.788	0.714	0.410	0.723	0.698	1.000	101	55	73	30	0	0	12
Intact.KEGGpath-Deg	0.755	0.783	0.728	0.441	0.694	0.682	1.000	134	55	93	40	0	0	2
Intact.KEGG2path-Deg	0.755	0.783	0.727	0.444	0.694	0.679	1.000	134	56	93	39	0	0	2
Union.Palsson_HIPOT_1-Trav	0.747	0.781	0.713	0.435	0.733	0.657	1.000	101	54	74	29	0	0	9
C														
Yipd.KEGG-SD	0.753	0.778	0.729	0.434	0.724	0.666	1.000	134	59	97	38	0	0	0
Yipd.KEGG2-SD	0.753	0.777	0.729	0.434	0.724	0.665	1.000	134	59	97	38	0	0	0
Yipd.KEGG-SNN	0.749	0.774	0.725	0.435	0.754	0.642	1.000	134	59	101	42	0	0	0
Yipd.KEGG2-SNN	0.749	0.773	0.725	0.437	0.716	0.667	1.000	134	59	96	37	0	0	0
Yipd.KEGGpath-SD	0.735	0.760	0.709	0.467	0.806	0.576	1.000	134	52	108	56	0	0	0
Yipd.KEGG2path-SD	0.734	0.759	0.709	0.465	0.813	0.574	1.000	134	53	109	56	0	0	0
Yipd.KEGGpath-SNN	0.730	0.756	0.704	0.472	0.716	0.623	1.000	134	53	96	43	0	0	0
Yipd.Palsson_HIPOT_4-CC	0.690	0.740	0.640	0.490	0.530	0.862	1.000	100	50	53	3	0	0	0
Yipd.Palsson_HIPOT_1-Deg	0.686	0.737	0.636	0.451	0.604	0.785	1.000	101	49	61	12	0	0	0
Yipd.Palsson_HIPOT_0-Deg	0.684	0.734	0.634	0.451	0.608	0.778	1.000	102	50	62	12	0	0	0
Yipd.Palsson_nonHIPOT_1-CC	0.673	0.725	0.621	0.538	0.480	0.865	1.000	98	43	47	4	0	0	0
Yipd.Palsson_nonHIPOT_2-CC	0.663	0.715	0.611	0.552	0.464	0.868	1.000	97	42	45	3	0	0	0
Yipd.Palsson_nonHIPOT_1-CCinv	0.662	0.713	0.612	0.540	0.480	0.856	1.000	98	44	47	3	0	0	0
Yipd.Palsson_nonHIPOT_3-CC	0.662	0.714	0.609	0.557	0.458	0.872	1.000	96	41	44	3	0	0	0
Yipd.Palsson_nonHIPOT_3-CCinv	0.657	0.709	0.605	0.559	0.458	0.861	1.000	96	41	44	3	0	0	0
Ypi.Palsson_HIPOT_2-Trav	0.653	0.694	0.613	0.509	0.653	0.627	1.000	101	55	66	11	0	0	0
Yipd.KEGGtype-SD	0.653	0.691	0.615	0.527	0.583	0.677	1.000	132	48	77	29	0	0	0
Ypi.Palsson_HIPOT_3-Trav	0.653	0.693	0.613	0.504	0.670	0.619	1.000	100	55	67	12	0	0	0
Yipd.KEGG2type-SD	0.652	0.690	0.614	0.528	0.576	0.686	1.000	132	45	76	31	0	0	0
D														
Intact.Palsson_nonHIPOT_2-Eccinv	0.563	0.608	0.518	0.826	0.938	0.176	1.000	97	24	91	67	0	0	0
Intact.Palsson_nonHIPOT_3-Eccinv	0.563	0.608	0.518	0.826	0.938	0.177	1.000	96	23	90	67	0	0	0
Intact.Palsson_nonHIPOT_4-Eccinv	0.563	0.608	0.518	0.826	0.938	0.177	1.000	96	23	90	67	0	0	0
Intact.Palsson_HIPOT_4-Eccinv	0.560	0.605	0.515	0.780	0.880	0.229	1.000	100	22	88	66	0	0	0
Union.KEGG2typepath-Eccinv	0.540	0.578	0.503	0.895	0.992	0.105	1.000	130	88	129	41	0	0	0
Union.KEGG2type-Eccinv	0.540	0.577	0.503	0.895	0.992	0.105	1.000	132	70	131	61	0	0	0
Union.KEGG2-Eccinv	0.540	0.577	0.503	0.895	0.993	0.105	1.000	134	80	133	53	0	0	0
Union.KEGG2path-Eccinv	0.540	0.577	0.503	0.895	0.993	0.105	1.000	134	49	133	84	0	0	0
Union.KEGGtypepath-Eccinv	0.540	0.577	0.503	0.894	0.992	0.106	1.000	130	83	129	46	0	0	0
Union.KEGGtype-Eccinv	0.540	0.577	0.503	0.894	0.992	0.107	1.000	132	70	131	61	0	0	0
Union.KEGG-Eccinv	0.540	0.576	0.503	0.893	0.993	0.107	1.000	134	80	133	53	0	0	0
Union.KEGGpath-Eccinv	0.540	0.576	0.503	0.894	0.993	0.106	1.000	134	95	133	38	0	0	0

Se muestran distintos pares red-centralidad y sus valores de **ABC** con los límites superior e inferior del intervalo de confianza de 99% (**99%ls** y **99%li** respectivamente) además de los valores de mínimo **error**, y de máximas **sensibilidad** y **especificidad**. El valor del índice de clasificación por incremento de centralidad se muestran en la columna con el título **CIC**. También se presentan el número total de **GICs** por red y cuántos de estos fueron correctamente clasificados con la RIM (**RIMVPs**) y la centralidad en cuestión, o cuántos fueron predichos con la RIDM y la centralidad en cuestión (**RIDMVPs**). También se muestran cuántos de estos RIDMVPs obtuvieron valores de DC menores, mayores o iguales a 0 (**DC<0**, **DC>0**, **DC=0** respectivamente). En la última columna se muestra el número de genes presentes en el conjunto diferencia entre RIMTPs y RIDMTPs, es decir aquellos que se clasificaban con la RIM pero ya no se pudieron clasificar con la RIDM.

III.c ALGUNOS GICs PRODUCEN PROTEÍNAS MULTIFUNCIONALES

Lo anterior pone en consideración por lo menos dos grupos de GICs: en un primer grupo se encuentran aquellos que pueden ser identificados tanto con la RIM como con la RIDM optimizada cuya predicción debería depender de sus relaciones químico-genéticas y ser robusta ante la adición de nuevas interacciones; en un segundo grupo se encuentran los GICs que solamente pueden ser clasificados usando la RIDM mixta donde la actividad enzimática no fue suficiente para explicar su indispensabilidad sugiriendo que presentan una función adicional. Así, una primera aproximación para saber si algún gen puede estar asociado a dos o más funciones, como en este caso, consistió en analizar los GICs en función de la posibilidad de que codifiquen proteínas con más de un dominio funcional (considerado como una región estructural asociada con una función particular) usando el método descrito en la sección **II.e**. En la **Tabla 8** se muestran los 134 GICs descritos por el SGDP que están contenidos en al menos una de las RIMs y cuyo contenido de dominios funcionales fue analizado a partir de la información reportada en la base de datos PFAM (<http://pfam.sanger.ac.uk/>).

En la **Tabla 8** también se indican con un símbolo > los 56 GICs' que fueron predichos exclusivamente al añadir interacciones que además aumentaron su centralidad en la RIDM Yipd.KEGG2path. Estos GICs' fueron seleccionados considerando que la centralidad SD sobre esta red Yipd.KEGG2path generó un $ABC = 0.734 \pm 0.025$ y un $CIC = 1$ (asociando los GICs predichos con por lo menos dos funciones distintas al clasificarlos por aumento de centralidad debido a interacciones de RIDs sobre RIMs; ver sección anterior) y, si bien éste no es el par red-centralidad con el que se obtuvo el mayor ABC con $CIC = 1$, es con el que se clasifica el mayor número de GICs a partir de la información de la base de datos KEGG (109 de 134 totales; ver **Tabla 1** y **Tabla 7**). Lo anterior permite notar que 25 de los 56 GICs' asociados a más de una función presentan más de un dominio mientras 31 de estos genes presentan un solo dominio. Estos últimos 31 genes representan más de la mitad de los GICs' predichos (55.3% de 56 totales) y plantean la existencia de GICs que se expresan como proteínas multifuncionales con un solo dominio. Esto muestra que la simple búsqueda de dominios PFAM no es suficiente evidencia de una o varias funciones.

Como una aproximación para buscar proteínas previamente reportadas como multifuncionales se realizó una búsqueda de proteínas tipo "moonlighting" mediante técnicas de minería de textos descritas en la sección **II.f**. En la **Tabla 9** se presenta una lista curada manualmente a partir del resultado obtenido por la minería de textos y se muestran los genes de *S. cerevisiae* que producen proteínas "moonlighting". Al generar esta tabla se observó que muchas de las funciones que presentan este tipo de proteínas están relacionadas con alguna actividad catalítica o metabólica y la mayoría de las funciones alternativas implican interacción con ácidos nucleicos y, algunas otras participan en la interacción con otras proteínas aisladas o en complejos multiprotéicos. De manera importante, solo los genes YKL060C y YLR355C de esta lista son indispensables para el crecimiento y se indican también en la **Tabla 8**.

El texto de los títulos y resúmenes correspondientes a los artículos que hacen referencia a enzimas y al término "moonlighting" fue obtenido y analizado mediante el método descrito en la sección **II.f**. Con ello se encontraron 655 objetos tipo proteína, de los cuales 10 se identificaron en *S. cerevisiae* y, entre ellos se detectaron los genes YBR020W y YDL130W (**Tabla 9**). Tanto los resultados anteriores del análisis de dominios PFAM como los obtenidos por minería de textos muestran que no es trivial relacionar un gen con proteínas que llevan a cabo distintas funciones y, es por ello que en la sección siguiente se presenta una estrategia preliminar para tratar de establecer dicha relación.

Tabla 8

ORF	Dominios PFAM	Descriptores UNIPROT	Nombre(s) de dominio PFAM
> YAR019C	1	CDC15	Pkinase
- YBR002C	1	RER2;YBR0107	Prenyltransf
> YBR029C	1	CDS1;CDG1;YBR0313	CTP_transf_1
- YBR153W	1	RIB7;YBR1203	RibD_C
- YBR196C	1	PGI1;YBR1406	PGI
> YBR252W	1	DUT1;YBR1705	dUTPase
> YBR256C	1	RIB5;YBR1724	Lum_binding
> YBR265W	1	TSC10;YBR1734	adh_short
> YCR012W	1	PGK1;YCR12W	PGK
> YDL045C	1	FAD1;D2702	PAPS_reduct
- YDL103C	1	QR11;UAP1;D2362	UDPGP
> YDL150W	1	RPC4;RPC53;D1557	RNA_pol_Rpc4
> YDR044W	1	HEM13;YD5112.02	Coprogen_oxidas
- YDR047W	1	HEM12;HEM6;POP3;YD9609.03	URO-D
> YDR050C	1	TP11;YD9609.05C	TIM
> YDR062W	1	LCB2;SCS1;TSC1;YD9609.16;D4246	Aminotran_1_2
- YDR208W	1	MSS4;YD8142A.05	PIP5K
> YDR236C	1	FMN1;YD8419.03C	Flavokinase
> YDR454C	1	GUK1;D9461.39	Guanylate_kin
> YER003C	1	PMI40	PMI_tysel
> YER023W	1	PRO3;ORE2	F420_oxidored
> YFL017C	1	GNA1;PAT1	Acetyltransf_1
> YFL022C	1	FRS2	tRNA-synt_2d
> YFL045C	1	SEC53;ALG4	PMM
- YGL001C	1	ERG26	3Beta_HSD
> YGL040C	1	HEM2	ALAD
- YGL155W	1	CDC43;CAL1;G1864	Prenyltrans
- YGR175C	1	ERG1	SE
> YGR185C	1	TYS1;MGM104;G7522	tRNA-synt_1b
> YGR267C	1	FOL2;G9349	GTP_cyclohydrol
> YHR143W-A	1	RPC10;RPB12;YHR143BW	DNA_RNApol_7kD
> YHR190W	1	ERG9	SQS_PSY
> YIR008C	1	PR11;YIB8C	DNA_primase_S
> YJL026W	1	RNR2;CRT6;J1271	Ribonuc_red_sm
> YJL090C	1	DPB11;J0918	BRCT
- YJL167W	1	ERG20;BOT3;FDS1;FPP1;J0525	polyprenyl_synt
> YJR006W	1	POL31;HUS2;HYS2;SDP5;J1427;YJR83.7	DNA_pol_E_B
> YJR016C	1	ILV3;J1450	ILVD_EDD
> YJR057W	1	CDC8;J1715	Thymidylate_kin
> YKL035W	1	UGP1;YKL248	UDPGP
> YKL045W	1	PRI2;YKL258	DNA_primase_lrg
> YKL060C	1	FBA1;YKL320	F_bP_aldolase
> YKL141W	1	SDH3;CYB3;YKL4	Sdh_cyt
> YKL152C	1	GPM1;GPM;YKL607	His_Phos_1
> YKL192C	1	ACP1	PP-binding
> YLR066W	1	SPC3;L2186	SPC22
> YMR113W	1	FOL3;YM9718.12	Mur_ligase_M
- YMR208W	1	ERG12;RAR1;YM8261.02	GHMP_kinases_N
> YMR296C	1	LCB1;END8;TSC2	Aminotran_1_2
> YNL113W	1	RPC19;N1937	RNA_pol_L_2
> YNL151C	1	RPC31;RPC8;ACP2;N1769	RNA_pol_3_Rpc31
> YNL247W	1	YNL247W;N0885	tRNA-synt_1e
> YNR003C	1	RPC34;N2031	RNA_pol_Rpc34
- YNR043W	1	MVD1;ERG19;MPD;N3427	GHMP_kinases_N
- YOL005C	1	RPB11	RNA_pol_L_2
- YOL097C	1	WRS1;HRE432	tRNA-synt_1b
> YOR074C	1	CDC21;CRT9;TMP1;YOR29-25	Thymidylat_synt
> YOR095C	1	RK11;YOR3174C	Rib_5-P_isom_A
> YOR176W	1	HEM15	Ferrochelataze
> YOR210W	1	RPB10	RNA_pol_N
> YOR224C	1	RPB8;YOR50-14	RNA_pol_Rpb8
> YOR236W	1	DFR1;O5231	DHFR_1
- YOR278W	1	HEM4;ORF1;O5463	HEM4
> YOR340C	1	RPA43;RRN12;O6271	SHS2_Rpb7-N
- YPL117C	1	IDI1;BOT2;LPH10C	NUDIX
> YPL204W	1	HRR25	Pkinase
> YPR113W	1	PIS1;PIS;P8283.5	CDP-OH_P_transf
> YPR175W	1	DPB2;P9705.7	DNA_pol_E_B
> YPR187W	1	RPB6;RPO26;P9677.8	RNA_pol_Rpb6
> YAL038W	2	CDC19;PYK1	PK;PK_C
> YBL035C	2	POL12;YBL0414	Pol_alpha_B_N;DNA_pol_E_B
> YBL076C	2	ILS1;YBL0734	Anticodon_1;tRNA-synt_1
> YBR038W	2	CHS2;YBR0407	Chitin_synt_1N;Chitin_synt_1
> YBR154C	2	RPB5;YBR1204	RNA_pol_Rpb5_N;RNA_pol_Rpb5_C
> YCL004W	2	PGS1;PEL1;YCL4W/YCL3W	PLDc_2;PLDc
> YDL055C	2	MPG1;PSA1;SRB1;VIG9	Hexapep;NTP_transferase

ORF	Dominios PFAM	Descriptores UNIPROT	Nombre(s) de dominio PFAM
- YDL205C	2	HEM3;D1057	Porphobil_deamC;Porphobil_deam
> YDR023W	2	SES1;SERS;YD9813.01	tRNA-synt_2b;Seryl_tRNA_N
> YDR037W	2	KRS1;GCD5;YD9673.09	tRNA-synt_2;tRNA_anti
YDR045C	2	RPC11;YD9609.01C	RNA_POL_M_15kD;TFIIS_C
> YDR353W	2	TRR1;D9476.5	Pyr_redox;Pyr_redox_2
YDR404C	2	RPB7;D9509.22	S1;SHS2_Rpb7-N
> YER043C	2	SAH1	AdoHcyase_NAD;AdoHcyase
YGL245W	2	GUS1;G0583;HRB724	tRNA-synt_1c_C;tRNA-synt_1c
> YGR094W	2	VAS1	Anticodon_1;tRNA-synt_1
> YGR264C	2	MES1	tRNA-synt_1g;MetRS-N
> YHR019C	2	DED81	tRNA-synt_2;tRNA_anti
- YHR072W	2	ERG7	Prenyltrans;Prenyltrans_2
YHR074W	2	QNS1	NAD_synthase;CN_hydrolase
YIL021W	2	RPB3	RNA_pol_A_bac;RNA_pol_L
YKL104C	2	GFA1;YKL457	SIS;GATase_2
YKL144C	2	RPC25;YKL1;UNF1	RNA_pol_Rbc25;SHS2_Rpb7-N
YLL018C	2	DPS1;APS1;APS;L1295	tRNA-synt_2;tRNA_anti
> YLR060W	2	FRS1;L2165	B5;B3_4
- YLR305C	2	STT4;L2142.4	PI3Ka;PI3_PI4_kinase
- YLR355C **	2	ILV5;L9638.7	IlvN;IlvC
YLR359W	2	ADE13;L8039.12	ADSL_C;Lyase_1
YML126C	2	ERG13;HMGS;YM4987.09C	HMG_CoA_synt_C;HMG_CoA_synt_N
- YMR220W	2	ERG8;YM9959.02	GHMP_kinases_C;GHMP_kinases_N
> YOR143C	2	THI80;YOR3373C	TPK_catalytic;TPK_B1_binding
> YPL028W	2	ERG10;LPB3	Thiolase_N;Thiolase_C
> YPL160W	2	CDC60;P2564	Anticodon_1;tRNA-synt_1
> YPR033C	2	HTS1;YP9367.13C	tRNA-synt_His;HGTP_anticodon
> YPR035W	2	GLN1;YP3085.01;YP9367.15	Gln-synt_C;Gln-synt_N
YPR110C	2	RPC5;RPC40;P8283.18	RNA_pol_A_bac;RNA_pol_L
YPR190C	2	RPC3;RPC82;P9677.11	RNA_pol_Rpc82;HTH_9
> YCL054W	2	SPB1;YCL54W;YCL431	Spb1_C;FtsJ;DUF3381
YDL102W	3	POL3;CDC2;TEX1;D2366	zf-C4pol;DNA_pol_B;DNA_pol_B_exo1
> YDR341C	3	YDR341C;D9651.10	tRNA-synt_1d;Arg_tRNA_synt_N;DALR_1
> YEL058W	3	PCM1;AGM1	PGM_PMM_I;PGM_PMM_IV;PGM_PMM_II
> YER171W	3	RAD3;REM1	DUF1227;Helicase_C_2;DEAD_2
> YER172C	3	BRR2;RSS1;SNU246;SYGP-ORF66	Sec63;Helicase_C;DEAD
- YHR020W	3	YHR020W	tRNA-synt_2b;ProRS-C_1;HGTP_anticodon
YLR153C	3	ACS2;L9634.10	AMP-binding;DUF3448;DUF4009
> YMR108W	3	ILV2;SMR1;YM9718.07	TPP_enzyme_M;TPP_enzyme_C;TPP_enzyme_N
- YNL256W	3	FOL1;N0848	FolB;Pterin_bind;HPPK
YNL262W	3	POL2;DUN2;N0825	DNA_pol_B;DNA_pol_B_exo1;DUF1744
- YNL267W	3	PIK1;N0795	PI3Ka;PI3_PI4_kinase;Pik1
> YOR335C	3	ALA1	DHHA1;tRNA-synt_2c;tRNA_SAD
> YIL078W	4	THS1	tRNA-synt_2b;TGS;HGTP_anticodon;tRNA_SAD
> YKL182W	4	FAS1	MaoC_dehydratas;Acyl_transf_1;zf-MaoC;DUF1729
YNL102W	4	POL1;CDC17;N2181	zf-DNA_Pol;DNA_pol_B;DNA_pol_B_exo1;DNA_pol_al pha_N
> YOR168W	4	GLN4;O3601	tRNA-synt_1c_C;tRNA-synt_1c;tRNA_synt_1c_R2;tRNA_synt_1c_R1
> YPL231W	4	FAS2;P1409	Ketoacyl-synt_C;adh_short_C2;ACPS;ketoacyl-synt Rapamycin_bind;PI3_PI4_kinase;DUF3385;FATC;FA T
- YKL203C	5	TOR2;DRR2;TSC14	RNA_pol_Rpb1_5;RNA_pol_Rpb1_3;RNA_pol_Rpb1_2;RNA_pol_Rpb1_4;RNA_pol_Rpb1_1
YOR116C	5	RPC1;RPO31;RPC160;O3254;YOR3254C	RNA_pol_Rpb1_5;RNA_pol_Rpb1_3;RNA_pol_Rpb1_2;RNA_pol_Rpb1_4;RNA_pol_Rpb1_1
YOR341W	5	RPA190;RPA1;RRN1;O6276	Biotin_carb_C;CPSase_L_D2;Carboxyl_trans;Biotin_I ipoyl;ACC_central;CPSase_L_chain
YNR016C	6	ACC1;ABP2;FAS3;MTR7;N3175	Ad_cyc_g-alpha;Guanylate_cyc;PP2C;LRR_8;LRR_1;LRR_4;R A
YJL005W	7	CYR1;CDC35;HSR1;SRA4;J1401	RNA_pol_Rpb2_3;RNA_pol_Rpb2_2;RNA_pol_Rpb2_5;RNA_pol_Rpb2_7;RNA_pol_Rpb2_4;RNA_pol_R pb2_1;RNA_pol_Rpb2_6
YOR151C	7	RPB2;RPO22;RPB150	RNA_pol_Rpb2_3;RNA_pol_Rpb2_2;RNA_pol_Rpb2_5;RNA_pol_Rpb2_7;RNA_pol_Rpb2_4;RNA_pol_R pb2_1;RNA_pol_Rpb2_6
YOR207C	7	RET1;RPC2;RPC128	RNA_pol_Rpb2_3;RNA_pol_Rpb2_2;RNA_pol_Rpb2_5;RNA_pol_Rpb2_7;RNA_pol_Rpb2_4;RNA_pol_R pb2_1;RNA_pol_Rpb2_6
YPR010C	7	RPA135;RPA2;SRP3;RRN2;YP9531.03C	RNA_pol_Rpb2_3;RNA_pol_Rpb2_2;RNA_pol_Rpb2_5;RNA_pol_Rpb2_7;RNA_pol_Rpb2_1;RNA_pol_R pa2_4;RNA_pol_Rpb2_6
YDL140C	8	RPO21;RPB1;RPB220;SUA8;D2150	RNA_pol_Rpb1_7;RNA_pol_Rpb1_5;RNA_pol_Rpb1_3;RNA_pol_Rpb1_R;RNA_pol_Rpb1_6;RNA_pol_R pb1_2;RNA_pol_Rpb1_4;RNA_pol_Rpb1_1
YHR128W	No determinado	FUR1	No determinado

Genes indispensables para el crecimiento (páginas anteriores). Se muestran los genes indispensables para el crecimiento (GICs) descritos por el Saccharomyces Genome Deletion Project (Cherry, et al., 2012) contenidos en al menos una de las RIMs utilizadas en este estudio. La primera columna muestra los ORFs ordenados alfabéticamente. La segunda columna muestra el número de dominios reportados para la proteína correspondiente según PFAM (Punta, et al., 2012). La tercera columna muestra distintos descriptores para la proteína en cuestión anotados en Uniprot – Swiss-Prot Knowledge base (UniProt Consortium, 2012). La última columna muestra el identificador que corresponde a cada uno de los dominios PFAM enumerados en la primera columna. Los genes reportados previamente como "moonlighting" son indicados con asteriscos * (Lu, et al., 2001) ** (Zelenaya-Troitskaya, et al., 1995) (ver Tabla 9). Se indican con un > los GICs clasificados en la red Ypd.KEGG2path y no en la red KEGG2path usando la centralidad SD. Se indican con un - los GICs que no pudieron ser clasificados antes del punto de mínimo error con el par red centralidad anterior.

Tabla 9

ORF	Alias	Función 1	Función 2	Referencia
YBL022C	PIM1; LON; YBL0440	Proteasa	chaperona	(Suzuki, et al., 1997)
YMR089C	YTA12; RCA1; YM9582.14C	Proteasa	chaperona	
YER017C	AFG3; YTA10	Proteasa	chaperona	
YPR024W	YME1; YTA11; OSD1; YP9367.04	Proteasa	chaperona	
YDL130W	RPP1B; RPLA3; L12EIIIB; RPL44P	Proteína ribosomal	Transactivador	(Tchórzewski, et al., 1999)
YDL081C	RPP1A; RPLA1; L12EIIA; RPA1	Proteína ribosomal	Transactivador	
YLR340W	RPP0; RPLA0; RPA0; RPL10E; L10E; L8300.8	Proteína ribosomal	Transactivador	
YML007W	YAP1; SNQ3; PAR1; PDR4; YM9571.12	Sensor de stress oxidativo	Factor de transcripción	(Delaunay, et al., 2002)
YIR037W	HYR1; GPX3; ORP1	Sensor de stress oxidativo	Transactivador	
YPR080W	TEF1; P9513.7; eEF1A; EF-1alpha	Factor de elongación de la traducción	Complejo con ZPR1 YGR211W control del ciclo celular	(Ejiri, 2002)
YJR048W	CYC1; J1653	Citocromo C	Apoptosis	(Lim, et al., 2002)
YLR117C	CLF1; NTC77; SYF3; L2952	pre-mRNA splicing	Replicación del ADN	(Zhu, et al., 2002)
YOR259C	RPT4; SUG2; CRL13; PCS1	Componente de la subunidad 19S del proteasoma	Transactivador de GAL4	(Gonzalez, et al., 2002)
YGL048C	RPT6; SUG1; TBY1; TBPY; CIM3; CRL3	Componente de la subunidad 19S del proteasoma	Transactivador de GAL4	
YML028W	cTPxI; TPX1; ZRG14; TSA1; TSA	Peroxidasa	Chaperona	(Jang, et al., 2004)
YDR453C	cTPxII; TSA2; D9461.38			
YLR304C	ACO1; GLU1; L8003.22	Aconitasa	Mantenimiento del ADN mitocondrial	(Chen, et al., 2005)
YKL060C	FBA1; YKL320	Aldolasa	Ensamblaje de V-ATPasa	(Lu, et al., 2001)
YER069W	ARG5.6	Biosíntesis de Arginina	Regulación de la expresión génica	(Hall, et al., 2004)
YMR311C	GLC8; YM9924.03C	Activador de fosfatasa	Inhibidor de fosfatasa	(Tung, et al., 1995)
YLR355C	ILV5; L9638.7	Biosíntesis de aminoácidos	Estabilidad del ADN mitocondrial	(Zelenaya-Troitskaya, et al., 1995)
YBR020W	GAL1; YBR0302	Galactokinasa	Regulador de la transcripción	(Gancedo y Flores, 2008)
YGL253W	HXK2; HKB; HEX1; NRB486	Hexocinasa	Regulación catabólica	
YDR173C	ARG82; IPK2; ARGR3; YD9395.06C	Catabolismo de Arginina	Expresión de genes del metabolismo de fosfato	
YPL111W	CAR1; LPH15W	Biosíntesis de arginina.	Interacción con arginasa	
YNL229C	URE2; N1165	Activación de transcripción	Peroxidasa	
YHR174W	ENO2; ENOB	Enolasa	Transporte mitocondrial	
YGR222W	PET54; G8527	Activador de la traducción	Splicing mitocondrial	(Kaspar, et al., 2008)
YKR072C	SIS2; HAL3	Determinación de halotolerancia	fosfopantotenil-cistein-decarboxilasa con YKL088W	(Ruiz, et al., 2009)
YOR054C	VHS3; YOR29-05	Control del ciclo celular	Fosfopantotenil-cistein-decarboxilasa con YKL088W	
YEL009C	GCN4; ARG9; AAS3	Factor transcripcional	Ribonucleasa	(Nikolaev, et al., 2010)
YNL037C	IDH1; N2690	Isocitrato deshidrogenasa	Unión a mRNA mitocondrial	(Lin, et al., 2001)
YOR136W	IDH2; O3326; YOR3326W	Isocitrato deshidrogenasa		

Genes de *S. cerevisiae* curados manualmente a partir de 194 artículos obtenidos como resultado de una búsqueda de PubMed para encontrar aquellos que hacen referencia al término enzima y al término "moonlighting". Se muestra el ORF y los distintos Alias asociados a éste junto con las 2 diferentes funciones que los caracterizan como "moonlighting" en cada una de 18 referencias distintas.

III.d HACIA EL DESARROLLO DE UNA ESTRATEGIA *IN VIVO* PARA EXPLICAR LA INDISPENSABILIDAD DE LOS GICs MULTIFUNCIONALES

En este caso se eligió el GIC *YDR050C* para analizar la función biológica de este tipo de genes mediante una estrategia *in vivo* que pudiera ser utilizada posteriormente para validar la utilidad del método aquí descrito. Lo anterior debido a que se encontró entre los GICs asociados a más de una función (**Tabla 8**) obtenidos mediante los métodos descritos en el apartado de **Estrategias *in silico***. Así, se eligió este GIC para generar una proteína con la sustitución del ácido glutámico número 97 por un residuo de glutamina (E97Q) que debería anular su actividad catalítica manteniendo sus IPPs considerando que tal mutación que abate la actividad de la Triosa Fosfato Isomerasa (TPI) de *P. falciparum* (ver detalles en el apartado de **Estrategias *in vivo*** de la sección de métodos y ver secciones **IV.c** y **IV.d** para una discusión más a detalle). Por tanto, se obtuvo una cepa de *E. coli* con el plásmido MoBY_TPI1_wt que contiene la secuencia del gen *YDR050C*. Dicho vector fue purificado (carril 1 **Figura 5**) para después analizar su peso molecular linearizándolo con la enzima *Eco*NI obteniendo un fragmento menor a 9 kilobases (**kbs**) correspondiente a los 8,981 pares de bases (**pbs**) del plásmido (carril 2, **Figura 5**; ver sitio de corte y tamaño del plásmido en la **Figura 1**). También se analizó el patrón de restricción obtenido usando la enzima *Xba*I debido a los dos sitios de restricción predichos usando la secuencia del plásmido, uno en el ORF y otro en una región no codificante del plásmido (**Figura 1**). Este análisis mostró dos fragmentos: uno de 6,688 pbs y otro de 2,293 pbs (carril 3, **Figura 5**; mapa en la **Figura 1**) que corresponden con lo esperado para el plásmido MoBY_TPI1_wt con el ORF *YDR050C*. Adicionalmente se confirmó la presencia del gen *YDR050C* realizando una PCR con un par de oligonucleótidos diseñados para obtener homología con la región 5' no codificante de este gen y con una región localizada ~400 pbs dentro del gen (carril 3, **Figura 5**; mapa en la **Figura 1**).

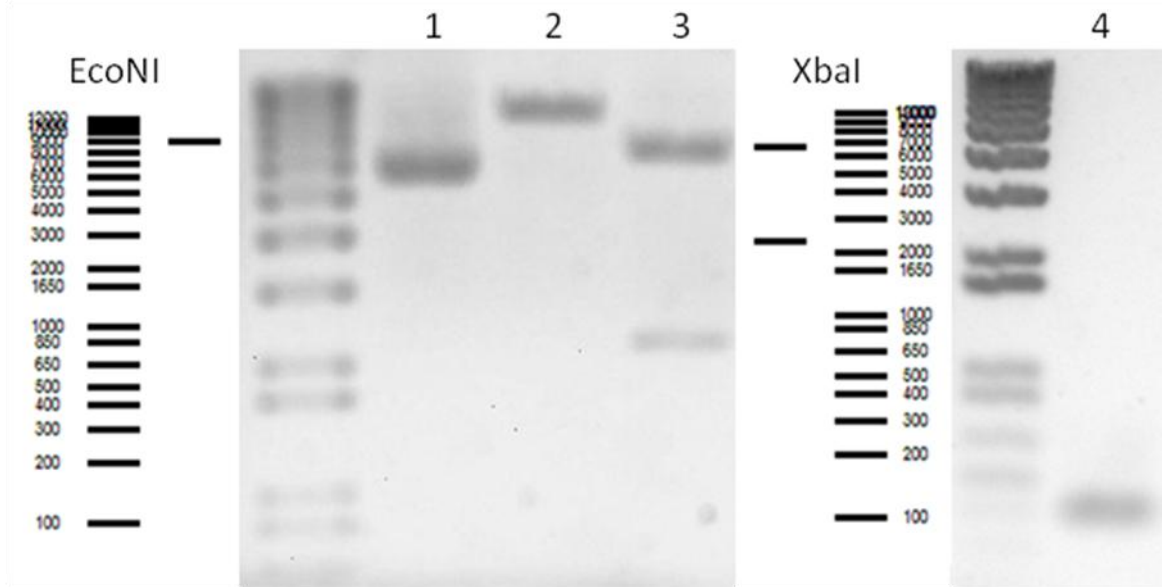


Figura 5 Verificación del plásmido. Se indican los patrones de restricción usando las enzimas *Eco*NI y *Xba*I. Se presentan diagramas con el peso molecular correspondiente a las bandas del marcador utilizado y el patrón de electroforesis predicho al utilizar la enzima correspondiente para digerir el plásmido MoBY_TPI1_wt. El **carril 1** corresponde al plásmido MoBY_TPI1_wt sin digerir, el **carril 2** presenta el resultado tras la digestión con *Eco*NI y el **carril 3** el patrón de digestión obtenido con la enzima *Xba*I. El gel en el extremo derecho muestra el **carril 4** con el producto de PCR obtenido al utilizar los oligos de confirmación de la **Tabla 5** cuyas posiciones se indican en la **Figura 1**.

A continuación se realizó la mutación E97Q mediante el procedimiento descrito en la sección **II.g** que funciona a partir de una reacción de PCR cuyo producto específico de amplificación se muestra en el panel superior izquierdo de la **Figura 6**. Estos productos de PCR fueron utilizados para transformar la cepa BW25142 (**Tabla 4**) y recuperar las células con el plásmido con la mutación al sembrarlas en los medios descritos en la sección **II.j**. Con ello se obtuvieron varias colonias transformantes resistentes a kanamicina y cloranfenicol que se muestran en los paneles superiores derecho y central de la **Figura 6** lo que confirmó la presencia del plásmido ahora llamado MoBY_TPI1_E97Q. Tres de estas colonias fueron utilizadas para purificar el plásmido con el gen *YDR050C* mutante y analizarlo por restricción como en el caso del plásmido con el gen silvestre (**Figura 6**, panel inferior).

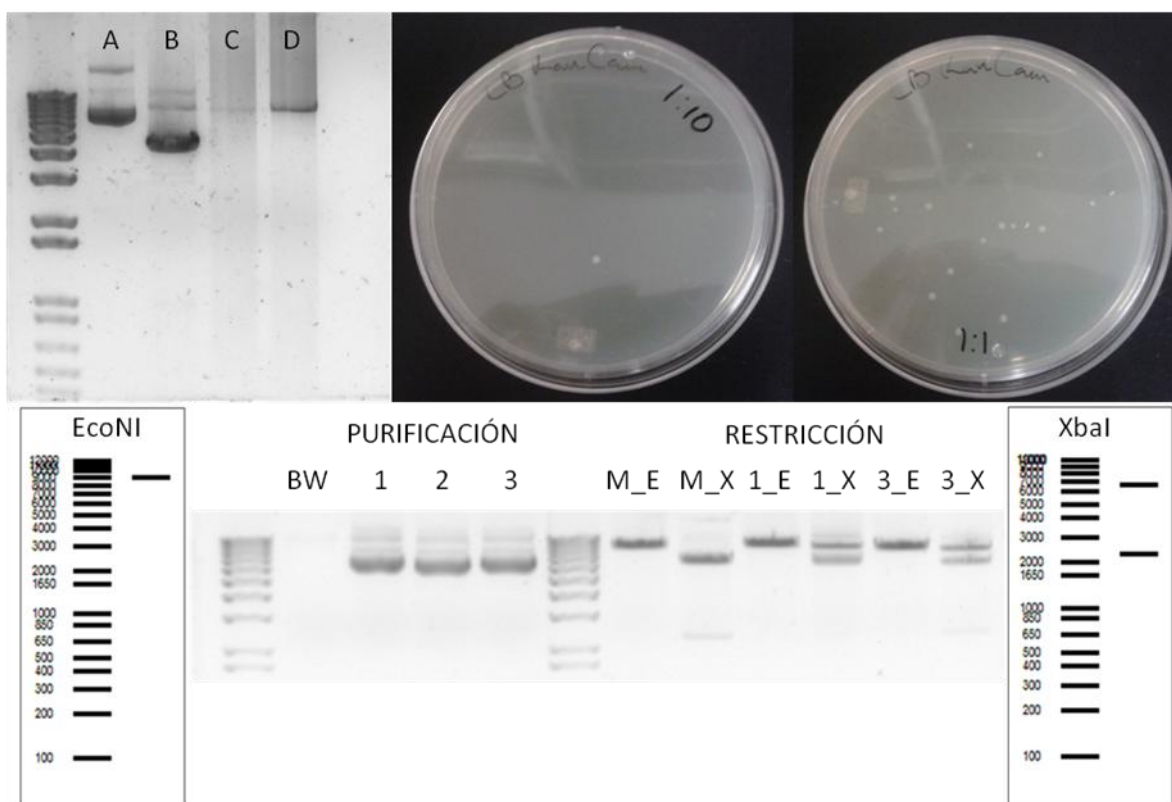


Figura 6 Mutagénesis. En la parte **superior izquierda** se muestra el resultado de la reacción de PCR mutagénica. El plásmido MoBY_TPI1_wt se presenta el **carril A**, el **carril B** muestra el control positivo mientras que los **carriles C y D** muestran los productos de la reacción de PCR utilizando 25 y 100ng del plásmido MoBY_TPI1_wt, respectivamente. En la parte **superior derecha y central** se muestran las colonias obtenidas al electrotransformar la cepa BW25142 con el producto de PCR mostrado en los **carriles C D** del panel anterior. En el **panel inferior** se muestra el ADN plasmídico purificado a partir de la colonia 1 (la única en la **placa superior central**) y las colonias 2 y 3 (obtenidas de la **placa superior derecha**). También se muestran los resultados obtenidos tras la digestión con las enzimas EcoNI y XbaI del plásmido MoBY_TPI1_wt (**carriles M_E y M_X** respectivamente) además de la digestión de los plásmidos obtenidos de las colonias 1 y 3 (**1_E, 1_X y 3_E con 3_X** igual que los anteriores). En los extremos del panel inferior se muestran los patrones de electroforésis predichos al utilizar cualquiera de las enzimas indicadas sobre cualquiera de los plásmidos en cuestión además de el peso molecular correspondiente a cada banda del marcador utilizado.

La secuencia de nucleótidos del ORF *YDR050C* en el plásmido MoBY_TPI1_E97Q se determinó a partir de un fragmento que contiene desde 137 pbs río arriba del gen hasta 136 pbs río abajo (**Figura 7**) detectando mutaciones en las posiciones 18, 289 y 360 del ORF correspondiente al gen *YDR050C*. Tanto la mutación que se encontró en el nucleótido 18 como la localizada en el 360 son silenciosas pues resultaron codificar para en el mismo aminoácido, mientras que la mutación del nucleótido 289 cambió el código para un ácido glutámico en la posición 97 por un residuo de glutamina. Esta última mutación solo se presentó en una de las dos colonias y el plásmido que la contiene se refiere como MoBY_ORF_E97Q.

```

ATG GCT AGA ACT TTC *TTC GTC GGT GGT AAC TTT AAA TTA AAC GGT TCC AAA CAA TCC ATT AAC GAA ATT GTT GAA AGA TTG AAC ACT GCT TCT ATC CCA G < 100
M A R T F F V G G N F K L N G S K Q S I K E I V E R L N T A S I P E
TAC CGA TCT TGA AAG AAgCAG CCA CCA TTG AAA TTT AAT TTG CCA AGG TTT GTT AGG TAA TTC CTT TAA CAA CTT TCT AAC TTG TGA CGA AGA TAG GGT C
10 20 30 40 50 60 70 80 90

AA AAT CTC GAA GTT GTT ATC TGT CCT CCA GCT ACC TAC TTA GAC TAC TCT GTC TCT TTG GTT AAG AAG CCA CAA GTC ACT GTC GGT GCT CAA AAC GCC TA < 200
N V E V V I C P P A T Y L D Y S V S L V K K P Q V T V G A Q N A Y
TT TTA CAG CTT CAA CAA TAG ACA CGA GGT CGA TGG ATC AAT CTG ATG AGA CAG AGA AAC CAA TTC TTC GGT GTT CAG TGA CAG CCA CGA GTT TTG CGG AT
110 120 130 140 150 160 170 180 190

C TTG AAG GCT TCT GGT GCT TTC ACC GGT GAA AAC TCC GTT GAC CAA ATC AAG GAT GTT GGT GCT AAG TGG GTT ATT TTG GGT CAC TCC cAAAGA AGA TCT < 300
L K A S G A F T G E N S V D Q I K D V G A K W V I L G H S Q R R S
G AAC TTC CGA AGA CCA CGA AAG TGG CCA CTT TTG AGG CAA CTG GTT TAG TTC CTA CAA CCA CGA TTC ACC CAA TAA AAC CCA GTC AGG gTTTCT TCT AGA
210 220 230 240 250 260 270 280 290

TAC TTC CAC GAA GAT GAC AAG TTC ATT GCT GAC AAG ACC AAG TTC GCT TTA GGT CAA CGcCTC GCT CTC ATC TTG TGT ATC GGT GAA ACT TTC GAA GAA A < 400
Y F H E D D K F I A D K T K F A L G Q C V G V I L C I G E T L E E K
ATG AAG GTC CTT CTA CTG TTC AAG TAA CGA CTG TTC TGG TTC AAG CGA AAT CCA GTT CCyCAG CCA CAG TAG AAC ACA TAG CCA CTT TGA AAC CTT CTT T
310 320 330 340 350 360 370 380 390

AG AAG CCC GGT AAG ACT TTC GAT GTT GTT GAA AGA CAA TTG AAC GCT GTC TTG GAA GAA GTT AAG GAC TGG ACT AAC GTC GTT GTC GCT TAC GAA CCA GT < 500
K A C K T L D V V E R Q L N A V L E E V K D W T N V V V A Y E P V
TC TTC CGC CCA TTC TGA AAC CTA CAA CAA CTT TCT GTT AAC TTG CGA CAG AAC CTT CTT CAA TTC CTG ACC TGA TTG CAG CAA CAG CGA ATC CTT GGT CA
410 420 430 440 450 460 470 480 490

C TGG GCC ATT GGT ACC GGT TTG GCT GCT ACT CCA GAA GAT GCT CAA GAT ATT CAC GCT TCC ATC AGA AAG TTC TTG GCT TCC AAG TTG GGT GAC AAG GCT < 600
W A I G T G L A A T P E D A Q D I H A S I R K F L A S K L G D K A
G ACC CGG TAA CCA TGG CCA AAC CGA CGA TGA GGT CTT CTA CGA GTT CTA TAA GTG CGA AGG TAG TCT TTC AAG AAC CGA AGG TTC AAC CCA CTG TTC CGA
510 520 530 540 550 560 570 580 590

GCC AGC GAA TTG AGA ATC TTA TAC GGT GGT TCC GCT AAC GGT AGC AAC GCC CTT ACC TTC AAG GAC AAG CCT GAT CTC GAT GGT TTC TTG GTC GGT GGT G < 700
A S E L R I L Y G G S A N G S N A V T F K D K A D V D G F L V G G A
CGG TCG CTT AAC TCT TAG AAT ATG CCA CCA AGG CGA TTG CCA TCG TTG CGG CAA TGG AAG TTC CTG TTC CGA CTA CAG CTA CCA AAG AAC CAG CCA CCA C
610 620 630 640 650 660 670 680 690

CT TCT TTG AAG CCA GAA TTT GTT GAT ATC ATC AAC TCT AGA AAC TAA < 747
S L K P E F V D I I N S R N *
GA AGA AAC TTC GGT CTT AAA CAA CTA TAG TAG TTG AGA TCT TTG ATT
710 720 730 740

```

Figura 7 Secuencia del gen *YDR050C* en el plásmido MoBY_ORF_E97Q. Se muestra la secuencia de nucleótidos del gen que codifica para la TPI1p en fragmento obtenido utilizando los oligos de secuenciación de la **Tabla 1** sobre el plásmido MoBY_ORF_E97Q. Las posiciones de estos oligos en dicho plásmido se indican en la **Figura 1**. Las mutaciones sinónimas encontradas se indican en rojo y la mutación 289 E97Q se indica en verde.

Para evaluar el fenotipo provocado por la mutación presente en el plásmido MoBY_TPI1_E97Q se utilizaron las cepas diploides heterocigotas Y26690 y Y23986 (**Tabla 4**) las cuáles se sometieron a los procedimientos de transformación, esporulación y selección resumidos en la **Figura 2** de la sección II.h. Tras recuperar varias colonias en los medios selectivos correspondientes, se utilizaron para confirmar su genotipo mediante PCR de colonia con los oligos de secuenciación de la **Tabla 5** en donde se esperaba amplificar solo el fragmento correspondiente al "cassette" de resistencia KanMX. Este "cassette" tiene un peso molecular mayor al del gen *YDR050C* por lo que se esperaría solamente la banda superior de las dos que se pueden observar en los carriles 5 y 13 del panel A de la **Figura 8** o en los carriles 3 y 5 del panel B de esta misma figura. Los resultados de este análisis muestran que la selección de haploides con canavanina resultó efectiva en la mayoría de los casos pero ninguna de las haploides seleccionadas presentaron solo el "cassette" KanMX4.

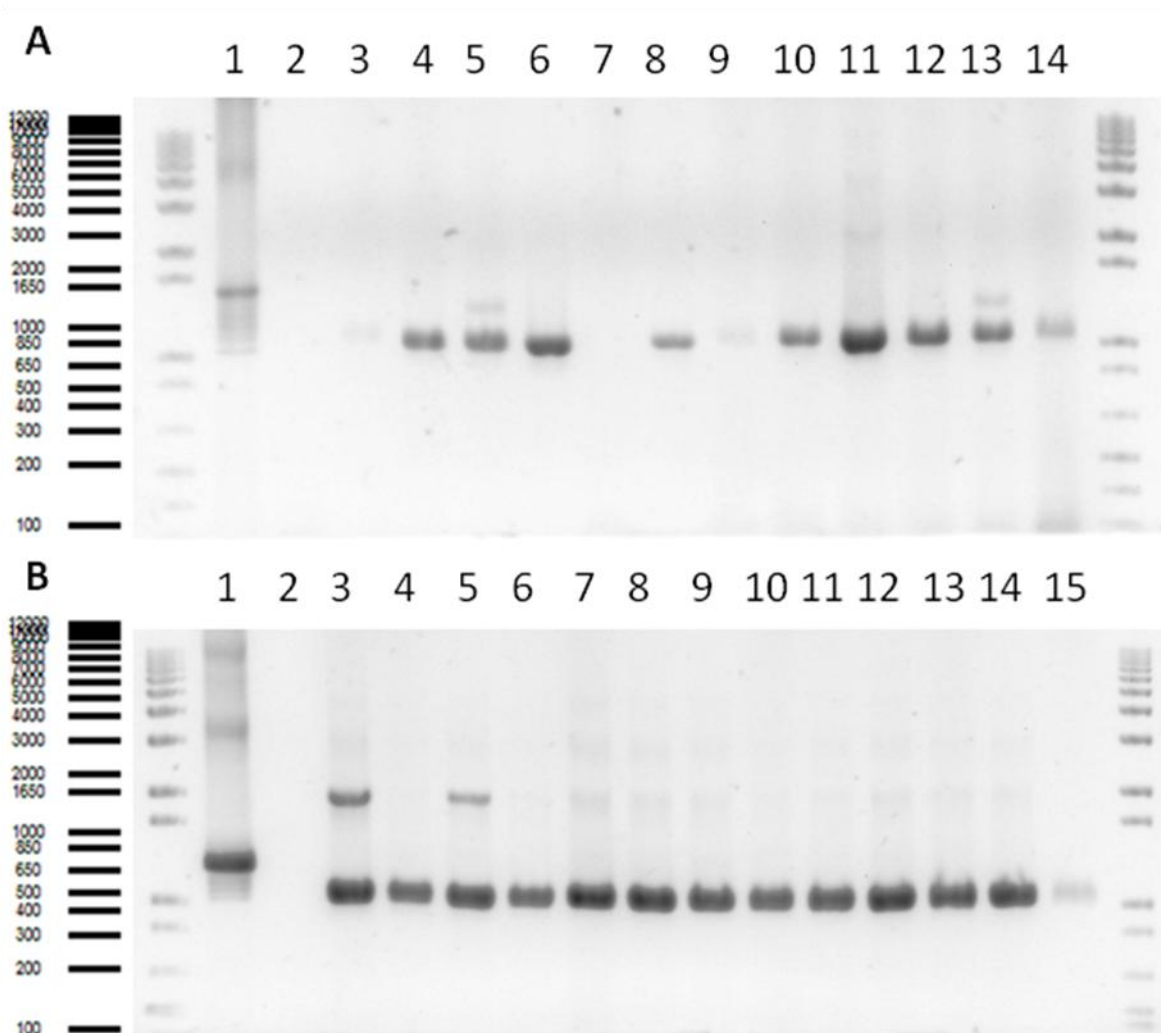


Figura 8 Selección de haploides por esporulación. Se muestran los productos de amplificación obtenidos al realizar PCRs con los oligos de secuenciación de la **Tabla 5** y el material genético de las distintas fuentes indicadas para cada carril. **Panel A:** 1)MoBY_TPI1_wt; 2)producto de reacción de PCR sin ADN; en los carriles 3-14 se muestran los productos obtenidos al realizar PCRs de colonia descritas en la sección II.h a partir de distintas variantes de la cepa diploide Y26690 que se indican a continuación: 3)sin modificación; 4)transformadas con MoBY_TPI1_wt; 5)transformadas con MoBY_TPI1_E97Q; los productos mostrados en los carriles 6-14 fueron obtenidos a partir de distintas colonias recuperadas en el medio de selección de esporas SC -arg suplementado con canavanina (60µg/ml), G418 (300µg/ml) e inositol (100µM) a partir de distintas variantes de la cepa diploide Y26690 como se indica a continuación: carriles 6-8 sin modificación; carriles 9-11 transformadas con el plásmido MoBY_TPI1_wt, carriles 12-14 transformadas con MoBY_TPI1_E97Q. **Panel B:** 1)plásmido MoBY_TPI1_wt; 2)producto de reacción de PCR sin ADN; los productos mostrados en los carriles 3-15 fueron obtenidos al realizar PCRs de colonia a partir de células de distintas cepas que se indican a continuación: 3)diploide Y26690; 4)haploide BY4741; los productos mostrados en los carriles 5-15 fueron obtenidos a partir de distintas colonias de células transformadas con el plásmido MoBY_TPI1_E97Q recuperadas en el mismo medio de selección de esporas que en los carriles 6-14 del **panel A** a partir de células de la cepa diploide Y26690.

Las cepas BY4741 y BY4742 de la **Tabla 4** también fueron utilizadas para probar una estrategia basada en el método de eliminación basado en PCR descrito en la sección II.j e intentaron recuperarse en los medios de cultivo YPED, YPEG y SC+ino descritos en la **Tabla 6** previamente utilizados para recuperar el crecimiento de células mutantes nulas para el gen *YDR050C* (**Tabla 6**). Para ello se generaron fragmentos de ADN con extremos homólogos al gen *YDR050C* o al gen *YFL026W* pero con el "cassette" KanMX4 en el lugar del ORF y tales fragmentos se utilizaron para transformar las cepas BY4741(MAT a) y BY4742(MAT α) mediante el método descrito en la sección II.i. Con lo anterior se verificó la falta de crecimiento en el medio de selección SC +ino con G418 donde fueron sembradas las células sometidas al proceso de transformación sin fragmento de ADN (**Figura 9**, panel inferior derecho).

Sin embargo, al utilizar el medio SC con sulfato de amonio y G418 se observó crecimiento incluso al sembrar las células sometidas al proceso de transformación sin ADN exógeno (**Figura 9** panel inferior izquierdo). Aún con lo anterior, las colonias recuperadas del panel superior derecho de la **Figura 9** se utilizaron para iniciar nuevos cultivos tanto en medio sólido como líquido pero no se detectó crecimiento significativo. Tales colonias también se utilizaron para hacer PCRs de confirmación pero de igual manera que en los casos anteriores tampoco se observó el genotipo esperado.

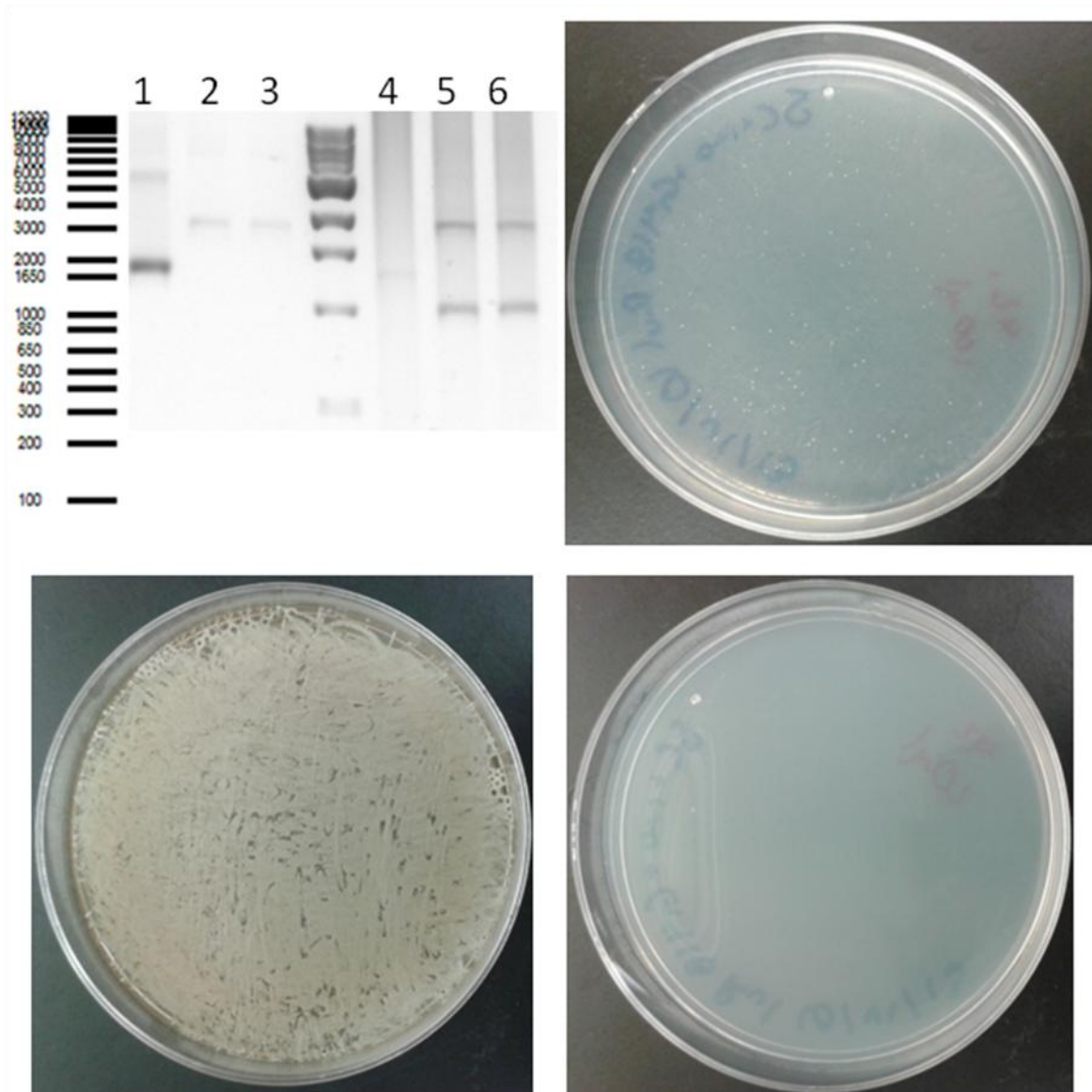


Figura 9 Delección por recombinación de productos de PCR. En el panel superior izquierdo se muestran los fragmentos de PCR que fueron utilizados. En el carril 1 se muestran los fragmentos amplificados utilizando los oligos UPSte2 y KanB (tabla) sobre la cepa DSte2 mientras que los carriles 2 y 3 muestran el resultado después de utilizar los oligos UPSte2 y DNSte2 sobre esta misma cepa. En el carril 4 se muestra el producto de PCR amplificado con los oligos UPseq y KanB sobre la cepa Δ TPI en los carriles 5 y 6 se muestran los 2 productos amplificados a partir de la cepa diploide Y26690. En el panel inferior izquierdo se muestra el crecimiento de las células sometidas al proceso de transformación sin ADN exógeno tras cultivarlas en medio SC con sulfato de amonio y G418. En el panel inferior derecho se puede verificar falta de crecimiento en las células sometidas al proceso de transformación sin fragmento de ADN sembradas en medio SC +ino con G418.

IV. DISCUSIÓN

IV.a ESTRUCTURA Y FUNCIÓN: REDES, CENTRALIDADES E INDISPENSABILIDAD

Una de las primeras y más importantes conexiones en torno a la relación estructura-función en redes biológicas fue realizada al ordenar todas las proteínas de la base de datos Dip (**Tabla 2**) según su número de IPPs buscando correlación con el efecto fenotípico de su remoción individual. De tal forma, las proteínas altamente conectadas dentro de una red de IPP (también llamadas "hubs") resultaron tener una probabilidad de ser indispensables para el crecimiento hasta tres veces mayor que las proteínas con pocas interacciones (Jeong 2001). Después se observó que al aumentar las conexiones de una proteína aumenta la probabilidad de que estén implicadas en más de una interacción hasta llegar el punto en que se vuelven indispensables (He y Zhang, 2006). También se ha mostrado que los genes humanos asociados con enfermedades que también son esenciales para el ratón tienden a estar más altamente conectados en redes de IPP y, al compararlos con respecto a otros genes asociados a enfermedades, muestran una mayor probabilidad de ser heredados de manera dominante (Skarnes, et al., 2011; Georgi, et al., 2013). Aparte de las redes de IPP, el análisis topológico de varias redes de interacciones genéticas también ha mostrado una correlación entre el grado y la indispensabilidad para el crecimiento de manera que eliminar los nodos más conectados tiene más impacto sobre la adecuación que la eliminación de nodos con pocas aristas (St Onge, et al., 2007). Además, los "hubs" de estas redes exhiben mayor pleiotropía, según lo estimado por la variedad de anotaciones funcionales, y el número de sus interacciones también ha sido correlacionado con su conservación en distintas especies de levadura (Costanzo, 2010).

Las causas y/o consecuencias de tales propiedades topológicas de las redes biológicas aún no se han establecido claramente pero ya se han utilizado de diferentes maneras. Un ejemplo es el caso en el que se determinaron las medidas más efectivas para identificar los genes indispensables para analizar estas mismas propiedades en las redes provenientes de organismos patógenos y utilizarlos como blancos farmacológicos (Estrada, 2006). En esta tesis, la razón principal tras la búsqueda de redes y centralidades que permitan distinguir los GICs consistió en acumular elementos para explicar su indispensabilidad. Una parte fundamental en este proceso recae en la asociación entre reacciones bioquímicas y los productos que las catalizan a partir de relaciones 'gen-proteína-reacción' (Palsson, 2009). Así, tanto la naturaleza de las interacciones con las que se construyeron las redes como la estructura que producen (reflejada en los valores de centralidad para cada nodo) permiten plantear hipótesis sobre qué es lo que determina que un gen sea, o no, indispensable para el crecimiento. Posteriormente tales hipótesis pueden ser verificadas o refutadas mediante diversos métodos experimentales.

Aquí se reconstruyeron redes a escala genómica a partir de reacciones enzimáticas de *S. cerevisiae* relacionando genes, indispensables o no, mediante diferentes tipos de interacciones, actividades o funciones biológicas. Estas relaciones funcionales diversas incluyen el total de interacciones reportadas en distintas bases de datos (**Tabla 2**) que fueron consideradas en ambas direcciones independientemente de en cuál estén descritas (detalles en la **sección IIa**). Tal consideración sobre la dirección de las interacciones cobra importancia, por ejemplo, en el caso de las bases de datos Yeastnet y Ypi donde ninguna de las interacciones está anotada en ambas direcciones. A esto se suma la observación de que en ninguno de los casos el número de interacciones reportadas en ambas direcciones representa al menos la quinta parte de las interacciones representadas en una sola dirección.

El método aquí utilizado ya había sido probado sobre RIMs mostrando que distintas medidas de centralidad fueron útiles para clasificar los GICs solamente cuando se usaron en forma combinada (del Río 2009). Lo anterior respondía al supuesto de que los GICs llevan a cabo funciones complejas que solo pueden ser abstraídas combinando distintas centralidades. Como alternativa, en este caso se consideró que las interacciones 'químico-genéticas' (o relaciones 'gen-proteína-reacción') incluidas en las RIMs no brindan una representación completa y se planteó la necesidad de incluir otras interacciones. Este

procedimiento que ahora considera nuevas interacciones en redes integradas resultó en 1311 pares RIDM-centralidad con $ABC > 0.5$ en contraste con las 14 de 198 posibilidades (7.07% del total) en las que se presentó un $ABC > 0.5$ usando RIMs (**Figura 4**). Con ello se corroboró que las RIDMs son significativamente más útiles para distinguir GICs que cualquiera de las RIMs o RIDs originales por separado.

Los resultados aquí presentados también se suman a los obtenidos en otros estudios que ya han confirmado esta relación que ha sido llamada 'centralidad-letalidad' en redes de IPP, no solo en levadura (Yu, et al., 2007; Zotenko, et al., 2008) sino en otros organismos (Hahn y Kern, 2005). Además, concuerdan con estudios anteriores que buscaron predecir y/o clasificar GICs utilizando conjuntos de datos que también empleados en este trabajo y que se analizaron mediante el mismo método de ABC por lo que es posible compararlos. Por ejemplo, al considerar características genómicas, filogenéticas y topológicas para predecir los genes descritos por el SGDP se obtuvo un $ABC = 0.82$ combinando el uso de métodos de inteligencia artificial sobre un conjunto balanceado de datos con número igual de genes indispensables y no indispensables (Saha y Heber, 2006). También se construyó una red para *S. cerevisiae* y, utilizando características topológicas, información de localización subcelular y procesos biológicos se obtuvo un $ABC = 0.808$ en el mejor de los casos (Acencio y Lemke, 2009). Ambos valores de ABC son superados por el que se obtuvo al usar el coeficiente de empacamiento (CC ver descripción en **Tabla 3**) sobre la red Yipd.KEGG2 (Yipd.KEGG2-CC $ABC = 0.896 \pm 0.024$ usando un intervalo de confianza del 99%; **Tabla 7 sección A**). Es importante destacar que el ABC reportado aquí se obtuvo usando hasta 134 GICs (**Tabla 8**) que participan en el metabolismo (debido a su participación en alguna de las 18 RIMs de la **Tabla 1**) mientras que el de los trabajos anteriores se obtuvo usando todos los GICs. Esto hace imprecisa la comparación pero es útil para contrastar los distintos enfoques. Por ejemplo, los estudios previos buscaron predecir GICs en forma confiable esperando facilitar su identificación y validar distintos conjuntos de GICs verificando los que ya están descritos y proponiendo nuevos. Por otra parte, aquí se plantea clasificar GICs de manera confiable para después analizar su función dentro del metabolismo analizando tanto las redes como las centralidades con las que fueron identificados.

Este análisis considera que las interacciones 'químico-genéticas' incluidas en las RIMs no brindan una representación completa de manera que es necesario incluir otras interacciones. Tal relación entre la esencialidad funcional y su papel dentro de varias redes biológicas (IPP, fosforilación, señalización, interacciones metabólicas, genéticas y regulatorias) integrándolas en una sola red global unificada ya fue cuestionada a partir del análisis de genes y sus roles tanto individuales como combinados (Khurana, et al., 2013). Así, se encontró que distintos parámetros mostraron una correlación positiva con el grado (centralidad deg **Tabla 3**) y solo fue significativa en las redes de IPP, de señalización y una red integrada que fue llamada "Multinet" en contraste con la red metabólica humana. Esto coincide con el estudio de Vitkup en el que encontró que las enzimas altamente conectadas tienen probabilidad menor o igual de ser indispensables en comparación con las enzimas menos conectadas en red metabólica de *S. cerevisiae* (Vitkup, et al., 2006) por lo que el grado no suele ser característico para los GICs en redes metabólicas como lo es para los GICs en redes de IPP (Jeong, et al., 2001).

Los resultados de esta tesis apoyan el supuesto de que al integrar distintos tipos de interacciones en una misma red es posible obtener una representación del metabolismo que permite la clasificación de GICs a partir de medidas estructurales. Para cumplir con lo anterior, la mejoría en la clasificación debería ser consecuencia del aumento en el valor de centralidad de los GICs contenidos en las distintas RIMs tras la adición de las interacciones de las diferentes RIDs en las nuevas RIDMs pero esto no se cumplió en todos los casos. Por ejemplo, el valor más alto de ABC obtenido al usar el coeficiente de empacamiento sobre la red Yipd.KEGG2 aumentó el ABC por 0.321 (Yipd.KEGG2-CC $ABC = 0.896 \pm 0.024$ **Tabla 7 sección A**; KEGG2-CC $ABC = 0.574 \pm 0.049$ ver Tablas C1 y C4 complementarias) identificando 49 GICs que no estaban clasificados con esta misma centralidad en la red KEGG2 (tomando como referencia el punto de mínimo error, ver sección **II.c**). No obstante, 24 de ellos no aumentaron su valor de centralidad tras la unión de las redes Yipd y KEGG2 además de que los 25 restantes no cambiaron su valor de centralidad (**Figura 10 Panel B**). Inclusive, entre los GICs predichos tanto con esta RIDM como con la RIM correspondiente se encontraron 39 sin modificación alguna en su valor para esta centralidad mientras que los 32 restantes presentaron una disminución (**Figura 10 Panel A**).

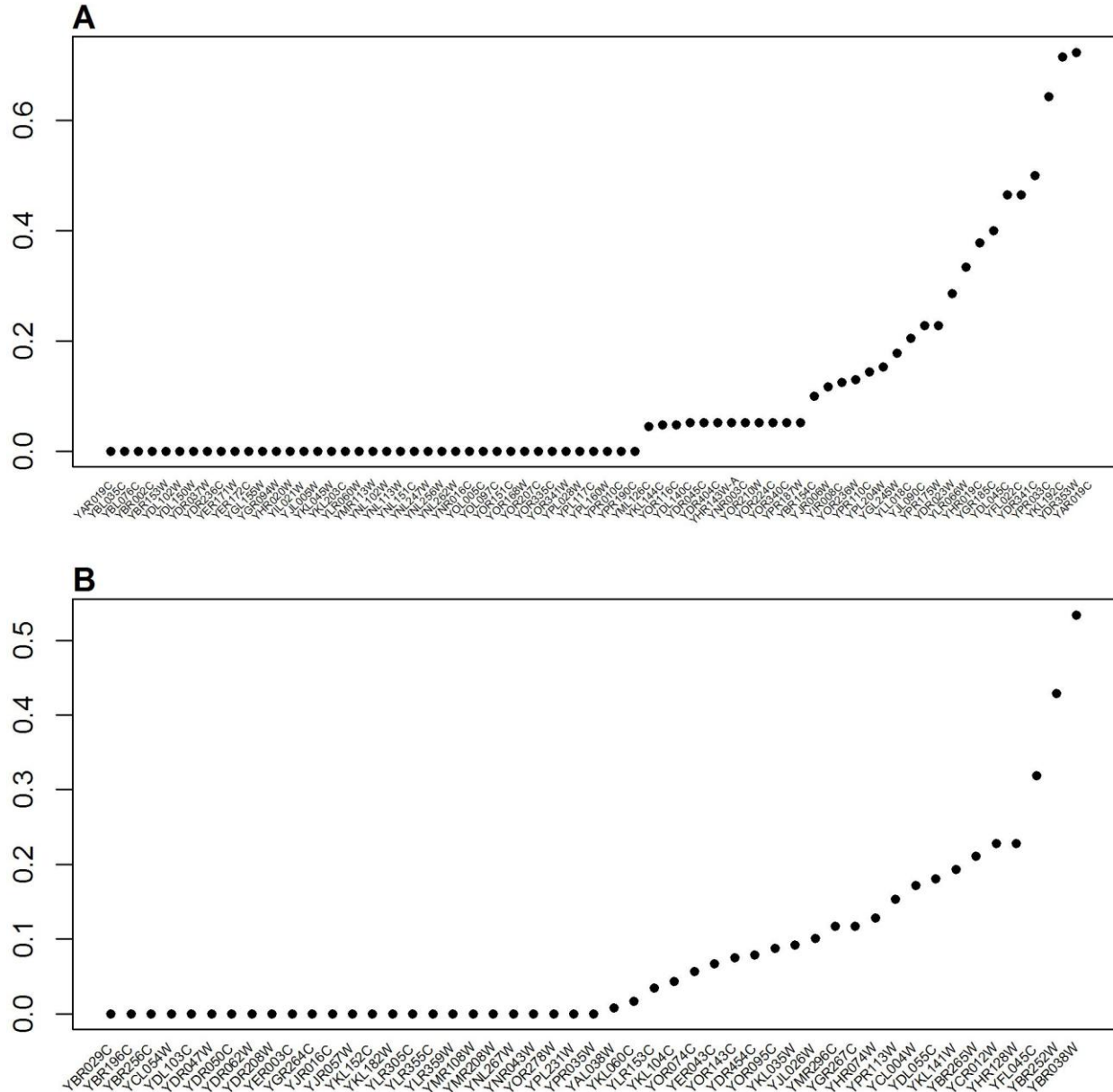


Figura 10 diferencias en el valor de coeficiente de empacamiento para los GICs correctamente clasificados. En el panel A se muestran los valores obtenidos al utilizar la centralidad CC sobre la red Yipd.KEGG2 para los 72 genes que fueron correctamente clasificados tanto con la red KEGG2 como con Yipd.KEGG2. En el panel B se muestran las diferencias entre los valores obtenidos para los 49 genes predichos exclusivamente en la red Yipd.KEGG2. Tales diferencias se calcularon restando el valor obtenido con la RIM menos el valor obtenido con la RIDM.

Estas observaciones señalan casos en los que la clasificación de GICs no está relacionada de manera directa con el incremento en su valor de centralidad tras la adición de interacciones exponiendo la necesidad de distinguir los GICs que aumentaron su valor de centralidad tras la adición de interacciones y, con ello, la probabilidad de clasificarlos. Como consecuencia se calculó la diferencia DC del valor de centralidad obtenido con la RIM original con respecto al valor obtenido con la misma centralidad sobre la RIDM correspondiente que, a su vez, se utilizó para calcular un índice de clasificación por incremento en centralidad (**CIC**) para cada red (ver sección II.d). Así, los pares red-centralidad con $ABC > 0.5$ y $CIC > 0$ son aquellos en donde la clasificación de GICs mejoró como consecuencia de incrementar la centralidad de los GICs al añadir interacciones en la red (**Tabla 7**,

sección C). De tal manera, los mejores de estos pares resultaron al clasificar los GICs según su grado esférico (o el número de nodos únicos que se encuentran a una distancia igual a 2 de cada nodo; centralidad SD **Tabla 3**) en las redes construidas a partir del conjunto unión de las IPP de la base de datos Yipd (**Tabla 2**) con la red tipo KEGG (**Tabla 1**). De este modo se encontraron 38 genes en la red Yipd.KEGG cuya clasificación solo fue posible tras considerar múltiples funciones que se reflejan en el aumento de su valor de la centralidad SD. Estos resultados sugieren que la actividad enzimática representada en las redes KEGG además de las IPP representadas en la red Yipd permiten explicar con una probabilidad de 73% la indispensabilidad de al menos el 28% de los GICs al clasificarlos según el número de nodos únicos que se encuentran a una distancia igual a 2 de cada nodo. Consecuentemente los 59 GICs restantes del total de 97 GICs fueron clasificados correctamente a partir de sus relaciones enzimáticas por sí mismas.

Esta clasificación de GICs por aumento de centralidad debida a la unión de interacciones de RIDs sobre RIMs y su relación con al menos dos funciones distintas no fue única. En este sentido, al usar la centralidad SD sobre la red Yipd.KEGG2path se obtuvo un $ABC = 0.734 \pm 0.025$ y un $CIC = 1$ (**Tabla 7**) y, si bien no es con el par red-centralidad con el que se obtuvo el mayor ABC con $CIC = 1$, es con este par con el que se clasificó el mayor número de GICs (109 de 134 totales **Tabla 7**). Tras lo anterior se encontró que en el 14% de los pares RIDM-centralidad la mejoría en la predicción de GICs fue debida al incremento en el valor de centralidad por la unión de RIMs y RIDs sin perder la capacidad predictiva ya se tenía en la RIM. Este efecto se presentó tras la integración de cualquiera de las RIMs con alguna de las RIDs Intact, Mpact, Yipd, Ypi y Union; todas ellas presentando principalmente interacciones proteína-proteína (IPP). Además, en este grupo aparecieron todas las centralidades locales a excepción de algunos casos en los que fueron utilizadas centralidades globales sugiriendo que son las IPP añadidas a estas redes las que permiten la clasificación de GICs mediante centralidades locales (**Tabla 7**). Estos resultados apoyan las observaciones de otros grupos (Pržulj, et al., 2004; Friedel y Zimmer, 2006; Hormozdiari, et al., 2007) que, al analizar redes de IPP, revelaron que los valores altos de grado no son la mejor medida para clasificar GICs.

IV.b IPP ENTRE ENZIMAS

Los GICs' predichos solamente tras el aumento en centralidad debido a la integración de IPP sobre casi cualquier RIM presentan una oportunidad interesante para evaluar la importancia de las interacciones físicas entre enzimas en relación con su indispensabilidad para el crecimiento. En este sentido, las IPP son un tema común al cuestionar la función de las proteínas. Por ejemplo, en algunos estudios se ha descrito que la interacción entre proteínas individuales puede resultar en el control fino de las tareas que lleva a cabo cada proteína además de evitar la agregación de proteínas o el despliegue de funciones promiscuas (Masino, et al., 2011; Pechmann, et al., 2009). Además, ya se ha discutido ampliamente que gran parte del trabajo dentro de las células es realizado por complejos compuestos por múltiples proteínas que están conectadas dentro de una misma unidad física (Alberts, 1998; Hartwell, et al., 1999). Tal concepto de complejo implica que cada proteína que lo compone puede participar en distintos complejos asociados a diferentes funciones, que la comunicación entre procesos implica IPP que conectan complejos y que las redes biológicas pueden exhibir propiedades emergentes que se comprenden mejor después de que sean descritas todas o, por lo menos la mayoría, de sus conexiones (Cusick, et al., 2005). Además, los complejos son relativamente independientes de manera que pueden ser reconstituidos en ausencia del resto de la red y trabajan juntos para realizar una misma tarea por lo que son tomados como ejemplos de módulos funcionales (Pereira-Leal, et al., 2006), tomando la definición de modularidad en el sentido de la teoría de redes (Newman, 2006).

La levadura *S. cerevisiae* ha probado ser un recurso extraordinario para tratar de entender la organización y comportamiento de estos complejos moleculares ya que existen al menos dos estudios globales dirigidos a la identificación de mapas comprensivos de los complejos proteicos de este organismo (Gavin, et al., 2006; Krogan, et al., 2006). Específicamente, se ha descrito que las proteínas producidas por GICs tienden a estar densamente conectadas dentro de complejos sugiriendo que la esencialidad es una propiedad modular más que una característica de proteínas individuales (Dezso, et

al., 2003; Hart, et al., 2007) de manera que ciertas proteínas son indispensables debido a su participación en módulos biológicos indispensables formados por complejos de proteínas indispensables densamente conectados (Zotenko, et al., 2008). Estas observaciones son congruentes con antecedentes en donde se detectó que los complejos más grandes tienen una mayor probabilidad de ser indispensables explicando por qué los GICs tienden a presentar un alto grado de interacción entre complejos (Wang, et al., 2009). También se ha encontrado que las proteínas indispensables tienden a presentar más interacciones dentro del mismo módulo funcional y que este grado de interacción intra-proceso correlaciona más con la esencialidad que el grado en general lo que demuestra que la esencialidad suele ser consecuencia de módulos indispensables que consisten de proteínas funcionalmente similares (Song y Singh, 2013). Esta esencialidad en complejos proteicos también ha sido observada en la levadura de fisión *Schizosaccharomyces pombe* (que es el único otro organismo con datos de esencialidad (Kim, et al., 2010)) sugiriendo que se trata de una característica general de distintas eucariotas ya que puede explicar diferencias en el estilo de vida de las dos levaduras (Ryan, et al., 2013).

Los resultados aquí presentados apoyan los modelos discutidos previamente en los que las proteínas son indispensables debido a su implicación en módulos funcionales. Acorde con ello, el mejor resultado fue obtenido la red Yipd.KEGG2, evaluada según el CC, mostrando que los nodos de la red con valores altos para esta medida de centralidad tienden a ser indispensables. Además, aún cuando esta clasificación no fue relacionada directamente con la adición de IPPs sobre las RIMs, las RIDMs construidas a partir de las mismas fuentes (Yipd y variaciones de KEGG) mantuvieron la alta probabilidad de clasificar GICs al ser ordenados según sus valores de SD y SNN. Estas tres medidas son locales en el sentido de que solo consideran las interacciones de cada nodo a uno y dos pasos de distancia dentro de la red y brindan cierta noción de modularidad debido a que los nodos con valores altos según estos índices presentan una densidad de conexión más alta que aquellos que tienen valores bajos. La diferencia entre CC con respecto a SD y SNN consiste en que la primera evalúa lo que podría considerarse como la formación de triángulos, dando valores más altos si ambos vecinos de un nodo están conectados entre sí, mientras que SD y SNN consideran las mismas interacciones a uno y dos pasos de cada nodo pero no toman en cuenta si sus vecinos también interactúan entre sí. De este modo, SD y SNN son más laxos que CC en cuanto a los nodos que presentan valores altos y reflejan la existencia de nodos con muchas interacciones que pueden ser mutuamente excluyentes. Lo anterior puede estar relacionado con estudios que han revelado que las IPP ocurren entre proteínas que pueden presentar tantas interfaces de unión como residuos expuestos en la superficie de manera que el carácter cooperativo o competitivo de estas interfaces puede ajustar la disponibilidad de proteínas dentro de la célula mediante el ensamblaje de complejos permanentes o transitorios que podrían funcionar conectando distintos módulos (Johnson y Hummer, 2013).

Desde un punto de vista evolutivo, se ha propuesto que los complejos multiprotéicos son favorecidos sobre proteínas individuales de gran tamaño (Lynch, 2011). Esto es debido a que las proteínas más grandes son difíciles de plegar y son más costosas de sintetizar mientras que las proteínas pequeñas que interactúan entre sí pueden plegarse independientemente en complejos más grandes. Más allá, se ha descrito que las proteínas de organismos más complejos están implicadas en más interacciones y forman complejos más grandes (comparados con sus formas primitivas) lo que podría interpretarse como un mecanismo para estabilizar proteínas que de otra manera serían monoméricas y susceptibles a fallas de plegamiento, desplegamiento u oligomerización tóxica (Fernández y Lynch, 2011). Por ello se ha propuesto hay proteínas que actúan como capacitores evolutivos al dar estabilidad adicional a cada proteína en la red de interacciones permitiendo que sus interactores exploren regiones menos estables mientras son estabilizadas (Rutherford, et al., 2007). De manera inversa, se esperaría que las proteínas inestables reciban estabilidad adicional de la red de interacciones (Dixit y Maslov, 2013). También se sabe que los cambios evolutivos en la estructura de proteínas tienen poca probabilidad de desestabilizar en gran medida el plegamiento nativo de una proteína esencial debido a que la pérdida completa de función no se puede sobrellevar y promueve el reclutamiento secundario de nuevas IPP que restablezcan la estabilidad estructural (Fernández y Lynch, 2011). Además, el estudio de Khurana y colaboradores (Khurana, et al., 2013) reveló que los genes altamente conectados en redes metabólicas tienden a presentar copias duplicadas planteando que, ante mutaciones de pérdida de función en alguna de estas enzimas, la vía metabólica en cuestión puede ser redirigida a rutas alternas que posiblemente implican copias duplicadas de la enzima desactivada.

Entonces, bajo la hipótesis de deriva génica, los organismos complejos desarrollan IPPs frecuentemente, no como vehículos inmediatos para funciones adaptativas, sino como mecanismos compensatorios para retener funciones clave (como es el caso de las proteínas producidas por los GICs) reduciendo así la necesidad de invocar ventajas selectivas directas a largo plazo (Fernández y Lynch, 2011)

Así, existe una aparente relación de las IPPs con el aumento de la capacidad de respuesta en la célula o como parte de mecanismos evolutivos para tolerar o evitar mutaciones dañinas. Debido a tal importancia, los GICs' predichos solamente tras la integración de las IPPs de Yípd sobre la red metabólica KEGG2path plantean una oportunidad interesante para preguntar si presentan por lo menos esas dos funciones, si pueden ser aisladas y si es alguna de ellas, o son ambas, lo que determina el fenotipo indispensable. Esto está relacionado con el entendimiento de la función biológica de las partes que componen distintas redes revelando, por ejemplo, la existencia de procesos de naturaleza multifuncional en los que uno o más componentes de una misma vía podrían actuar de diferente manera ante diversas circunstancias como se plantea a continuación.

IV.c ENZIMAS MULTIFUNCIONALES INDISPENSABLES PARA EL CRECIMIENTO

En las secciones anteriores se muestra que entre los GICs clasificados tanto con las RIDMs como con las RIDs se pueden encontrar genes cuya indispensabilidad podría radicar en su actividad enzimática aunque esto no significa que no puedan llevar a cabo más de una función. De hecho, estos GICs podrían realizar una o más funciones catalíticas distintas o incluso participar en interacciones físicas adicionales. En contraste, aquellos GICs' que son clasificados exclusivamente con las RIDMs relacionan su indispensabilidad con la presencia de más de una función aunque no dejan claro si ambas determinan esta indispensabilidad pero por lo menos una de las dos tiene que ser importante. De manera interesante, se ha observado la existencia de GICs asociados con complejos no indispensables que tienden a ser parte de un segundo complejo y que se podrían explicar porque la proteína en cuestión posee características estructurales únicas (actividad catalítica o dominio de reconocimiento estructural) requeridas por ambos complejos o bien porque tienen distintas características que se requieren en distintos complejos (Ryan, et al., 2013).

Cuando se consideran las interacciones reportadas en todas las bases de datos utilizadas en este estudio (**Tabla 1; Tabla 2**), las proteínas que participan en el metabolismo de *S. cerevisiae* tienen un promedio de 28.9 relaciones funcionales que representan múltiples actividades y señalan por lo menos dos grupos de GICs. En el primer grupo se encuentran aquellos que pueden ser identificados tanto con la red metabólica como con la red optimizada cuya predicción depende sus relaciones 'gen-proteína-reacción' y es robusta ante la adición de nuevas interacciones de otros tipos. El segundo grupo incluye los que solamente se pueden clasificar usando la red mixta donde la actividad enzimática no es suficiente para explicar su indispensabilidad sugiriendo que ejecutan múltiples funciones. Así, la capacidad de predecir GICs usando información 'químico-genética' relaciona la indispensabilidad de los genes metabólicos con la producción de proteínas con actividad enzimática. Sin embargo, el hecho de que las interacciones físicas mejoren dicha capacidad las añade como elementos que determinan la indispensabilidad de algunos de estos genes. Hasta este punto no queda claro si ambas funciones provocan este fenotipo o cuál de estas dos es la más importante y en adelante se argumenta al respecto.

Los GICs identificados entre los nodos con mayor valor de centralidad en una RIDM que no fueron clasificados correctamente en la RIM correspondiente indican que para capturar su naturaleza indispensable se requiere considerar relaciones funcionales adicionales a la actividad catalítica. Así, al combinar el ABC con el CIC se identificó que los pares red-centralidad con los que se obtuvieron los mejores resultados al clasificar los GICs provienen de las redes construidas a partir del conjunto unión de las interacciones en la base de datos Yípd con distintas modificaciones de las redes tipo KEGG2 (**Tabla 7**) evaluadas según su grado esférico (o el número de nodos únicos que se encuentran a una distancia igual a 2 de cada nodo; centralidad SD **Tabla 3**). De este modo se encontraron 38 genes cuya clasificación solo fue posible tras considerar múltiples funciones que se reflejan en un aumento de su valor de centralidad. Estos resultados sugieren que la actividad enzimática representada en las redes

KEGG además de las IPP representadas en la red Yipd permiten explicar con una probabilidad de 73% la indispensabilidad de al menos el 28% de los GICs al clasificarlos según la centralidad SD. Consecuentemente los 59 GICs restantes de un total de 97 GICs fueron clasificados correctamente como consecuencia de sus relaciones enzimáticas por sí mismas. Al usar esta centralidad sobre la red Yipd.KEGG2path se obtuvo un $ABC = 0.734 \pm 0.025$ y un $CIC = 1$ (asociando los GICs predichos con por lo menos dos funciones distintas al clasificarlos por aumento de centralidad debido a interacciones de RIDs sobre RIMs; **Tabla 7**) y si bien no es con el par red-centralidad con el que se obtuvo el mayor ABC con $CIC = 1$, es con este par con el que se clasifica el mayor número de GICs (109 de 134 totales **Tabla 7**).

Además, se sabe que las proteínas pueden presentar distintos grupos y dominios funcionales que las relacionan con distintas actividades. De tal manera, distintas regiones de una misma proteína podrían participar en una reacción enzimática mientras otras median la unión con otras proteínas e incluso algunas otras podrían fijar grupos prostéticos, todo esto de manera secuencial, excluyente o simultánea. Esto llevó a notar que 25 de los 56 GICs' asociados a más de una función y clasificados por aumento en centralidad presentan más de un dominio PFAM y los 31 restantes ostentan un solo dominio representando más de la mitad de los GICs' predichos (55.3% de 56 totales). Esto sugiere la existencia de GICs que se expresan como proteínas multifuncionales con un solo dominio lo cual no concuerda con la definición de dominio funcional (**Tabla 8**). Esta observación podría deberse a la existencia de algún tipo de presión evolutiva dentro de la célula relacionado al conflicto adaptativo que supone la existencia de proteínas multifuncionales con un solo dominio (Jeffery, 2004) pero hasta este punto no es posible explicar tal observación que resulta intrigante.

En este sentido, uno de los primeros descubrimientos relacionados con la multifuncionalidad de enzimas metabólicas fue reportado al demostrar mediante evidencias experimentales que, tanto en pato como en pollo, la proteína estructural δ -crystallin y la enzima metabólica argininosuccinato-liasa comparten el mismo gen por lo que dicho fenómeno fue referido como "gene sharing" (Piatigorsky, et al., 1988). Como resultado, la idea de un gen - una proteína - una función comenzó a parecer cada vez más simplista debido a que un número creciente de proteínas han demostrado tener diferentes funciones añadiendo una nueva dimensión a la complejidad celular. Fue hasta después que se comenzó a utilizar el término "moonlighting" para referirse a las proteínas multifuncionales excluyendo aquellas que son el resultado de fusiones génicas, proteínas homólogas no-idénticas, variantes de "splicing", proteínas con modificaciones post-traduccionales variables y aquellas que tienen una sola función que varía dependiendo de su localización o del sustrato que utilice (Jeffery, 1999). Adicionalmente, la diversidad de funciones no enzimáticas adicionales que exhiben estas enzimas incluye desde la transducción de señales hasta la regulación transcripcional y apoptosis pasando por crecimiento y motilidad e incluso funciones estructurales (Gancedo y Flores, 2008). Por ejemplo, un mapa metabólico de la glucólisis elaborado por el grupo de Sriram en el 2005, reveló que por lo menos 7 de las 10 enzimas glucolíticas y 7 de las 8 enzimas del ciclo de los ácidos tricarbónicos exhiben varias actividades en distintos organismos (Sriram, et al., 2005). Así, recientemente se ha reconocido que varias de las enzimas metabólicas que son "housekeeping" (Thellin, et al., 1999) participan en distintas funciones biológicas dentro y fuera de la célula debido a que se encuentran casi en cualquier circunstancia y en gran cantidad por lo que pueden adquirir nuevas funciones con mayor probabilidad (Reznik, et al., 2013).

Esto señala que esta multifuncionalidad podría ser solo la punta del iceberg y hasta el momento no se ha abordado el tema de la complejidad adicional que supone el hecho de que distintas redes bioquímicas se entrelacen funcionalmente mediante este mecanismo (Heinemann y Sauer, 2010). Así, existe cada vez más interés en el estudio de las proteínas denominadas "moonlighting" de manera que el número de publicaciones asociadas a este término ha estado en constante crecimiento desde que fue acuñado y un buen número de estas proteínas han sido reportadas en *S. cerevisiae* (Gancedo y Flores, 2008). Esta descripción aparentemente buena de las proteínas "moonlighting" presenta una oportunidad para detectar proteínas que han sido descritas como multifuncionales y buscar si se encuentran entre los GICs' descritos al considerar por lo menos dos funciones distintas. En este punto se podría encontrar utilidad adicional para el método aquí presentado debido a que relaciona la esencialidad de enzimas metabólicas con la realización de por lo menos una función adicional a su actividad catalítica y puede ser una herramienta para detectar y/o predecir proteínas con múltiples funciones relacionadas a un mismo papel fisiológico (representado por su indispensabilidad para el crecimiento y su implicación en el

metabolismo). En el caso específico de los GICs' aquí descritos se podría reflejar la multifuncionalidad de enzimas metabólicas a partir de IPP además de su actividad catalítica. Este grupo de genes indispensables multifuncionales incluye una porción considerable de las reacciones a lo largo de la glucólisis y, de manera interesante, se encuentran en los puntos donde se conecta esta vía con la biosíntesis de lípidos y de aminoácidos o con la fosforilación oxidativa. Otros de estos genes están dispersos en el metabolismo de nucleótidos y en el de aminoácidos y el de xenobióticos en menor proporción (**Figura 11**). Estos resultados coinciden con la propuesta en la que tener distintas funciones podría ser un importante mecanismo de regulación entre distintos procesos (Jeffery, 2003) que ya ha sido observado tanto en *S. cerevisiae* en otros organismos (Sriram, et al., 2005).

Ya se ha discutido que la pertenencia de proteínas indispensables en complejos no indispensables se puede explicar parcialmente por la presencia de proteínas multifuncionales cubriendo ~50% de los genes indispensables en complejos no indispensables (Zotenko, et al., 2008; Song y Singh, 2013; Ryan, et al., 2013). Esto puede ser debido a que estos genes son miembros de distintos complejos pero su indispensabilidad no ha sido determinada experimentalmente o documentada en bases de datos. Tal es el caso de SDH3 que esta anotada como parte del complejo de succinato deshidrogenasa y es el único miembro esencial de este complejo (Oyedotun y Lemire, 2004) por lo que se sugirió un papel adicional fuera de este que fue confirmado al identificarlo como miembro del complejo de inserción de la membrana interna mitocondrial TIM22 (Gebert, et al., 2011). En este sentido, el gen *YKL141W* relacionado con la producción de la enzima con actividad de SDH3 es predicho como GIC' en este estudio (**Tabla 8**). Esto significa que su indispensabilidad pudo ser relacionada con valores altos de centralidad solo cuando se consideraron las IPPs sobre sus relaciones enzimáticas. Más aún, la proporción de GICs' de este tipo cuya indispensabilidad se puede asociar con actividad catalítica e IPPs también es de ~50% (56 de 109 predichos con la centralidad SD en la red Yipd.KEGG2path) como encontraron Oyedotun y Lemire.

Más allá de los complejos multiprotéicos, se sabe que existen circuitos de regulación genética que son típicamente multifuncionales. Esto en relación a la observación de que, al incrementar el número de funciones posibles para cada gen, el número de genotipos posibles disminuye exponencialmente. Aún así, la cantidad de genes puede mantenerse relativamente grande cuando estos genes producen proteínas con un número modesto de funciones cuya adquisición depende de la función previa (Payne y Wagner, 2013). Así, dado que la caracterización en tiempo y forma de las funciones realizadas por las distintas biomoléculas en redes biológicas se relaciona con especulaciones en torno al control y la elucidación de principios de organización en sistemas biológicos (Heinemann y Sauer, 2010), la importancia de que una proteína presente múltiples funciones podría estar relacionada su habilidad para llevar a cabo operaciones de control; por ejemplo, utilizando una función como sensor de las condiciones celulares y una segunda función para actuar acorde al estímulo correspondiente (Ma, et al., 2010). Tales elementos de control pudieran jugar un papel indispensable en el funcionamiento o mantenimiento del metabolismo visto como sistema complejo y cada vez es más común encontrar que la multifuncionalidad es una estrategia general para aumentar el número de actividades atribuidas a una misma proteína sin aumentar el número de estas codificadas por el genoma. Lo anterior sugiere que mientras la(s) función(es) adicional(es) de una proteína no interfiera(n) con su función original la célula puede resultar beneficiada. Existen varias formas en que las proteínas multifuncionales pueden resultar en una ventaja competitiva y la más simple es que, al tenerlas, la célula necesitaría sintetizar menos proteínas y replicaría menos ADN. Además, pueden proveer estrategias para coordinar actividades celulares, regular vías catabólicas o biosintéticas en forma conjunta y disparar múltiples respuestas al mismo tiempo. También pueden proveer métodos para responder ante condiciones variables o ser parte de mecanismos de retroalimentación. Otro tema con impacto significativo sobre las redes metabólicas incluye la promiscuidad enzimática que puede llevar a conectividades más altas que la que se representa en las reconstrucciones actuales y es un tema muy real sobre todo en el metabolismo central. Dicha promiscuidad puede estar relacionada con la descripción de nuevas reacciones e incluso con la especificidad o flexibilidad en el uso de cofactores (Heinemann y Sauer, 2010).

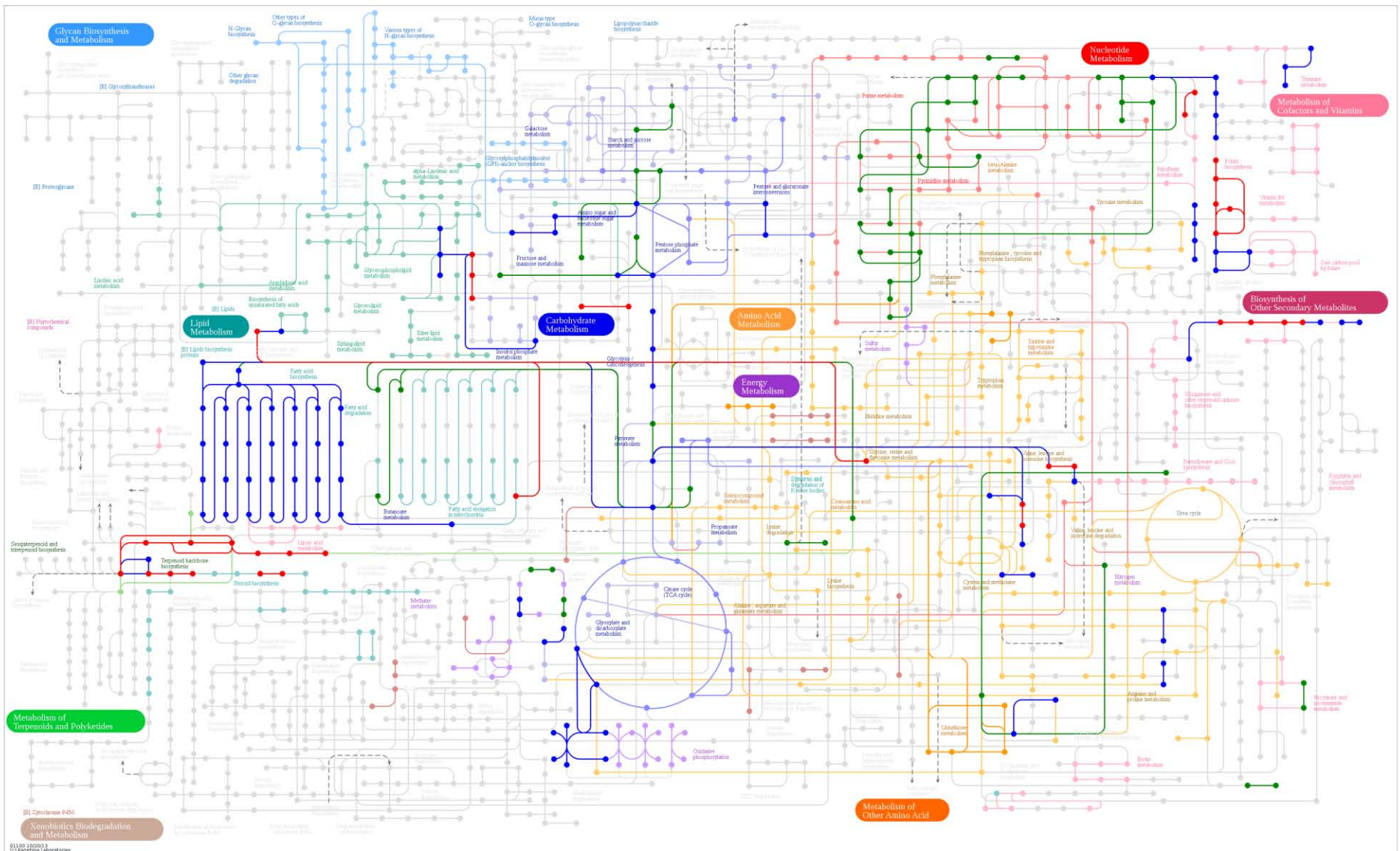


Figura 11 Mapa metabólico de *S. cerevisiae*. Se muestra un mapa con las reacciones anotadas en la base de datos KEGG, el color señala las vías metabólicas de *S. cerevisiae*: en verde se muestran las reacciones en las que están involucrados los GICs clasificados en la red KEGG2path usando la centralidad SD, en azul los GICs asociados con más de una función al clasificarlos con la misma centralidad en Ypd.KEGG2 con la misma centralidad.

En consecuencia, cuando se habla de múltiples funciones primero es necesario definir qué se entiende por función biológica y como se puede diferenciar entre ellas. Es por ello que la anotación funcional de proteínas es un problema abierto y constituye un reto principal para la biología molecular actual (Schnoes, et al., 2009). En este sentido, la función biológica se refleja en una actividad que puede determinarse mediante experimentos que comúnmente son dirigidos hacia esa actividad particular. Aun así, los análisis experimentales han servido para elaborar catálogos con anotación funcional para buscar que es lo que se sabe acerca de distintas proteínas para las que se ha observado cierta actividad específica. Esto ha resultado en extensas labores de anotación funcional escritas en lenguaje natural provocando descripciones complejas e intrincadas que varían entre distintos grupos acumulando un vocabulario en constante invención y re-inventación donde varios de los términos son sinónimos que no solo confunden a los curadores sino que además incrementa las probabilidades de errar en la interpretación (Janga, et al., 2010). En este sentido existe la base de datos UniprotGOA (Camon, et al., 2004) que es compilada a partir de las anotaciones realizadas por varios miembros del consorcio GO (Ashburner, et al., 2000; The Gene Ontology Consortium, 2004) representa el estado actual de la visión sobre el espacio de función de las proteínas y se ha considerado que las interacciones que comparten términos GO hasta el sexto nivel o más sugiere que las interacciones son reales (Zhang, et al., 2012). Esta ha mostrado que las anotaciones que provienen de experimentos de alto rendimiento son menos informativas que las que provienen de experimentos de menor escala además de que las anotaciones que provienen de experimentos de gran escala están sesgados hacia un número limitado de funciones. Además, muchos de estos experimentos se sobrelapan en las proteínas que anotan y en las anotaciones asignadas dando una idea del espacio de funciones proteicas y como toma forma a partir de la literatura afectando la habilidad de entender propiamente la función reportada con la función predicha y la función real (Schnoes, et al., 2013).

Como alternativa, han surgido métodos automáticos de inferencia de función como la identificación de proteínas multifuncionales abordada utilizando enfoques bioinformáticos (Tsai, et al., 2009; Han, et al., 2004), tomando en cuenta el conocimiento previo sobre regiones de unión que se sobreponen para predecir múltiples interacciones potenciales (Kim, et al., 2006) incluso cuando esta información no siempre se encuentra disponible (Ogmen, et al., 2005). También se ha intentado establecer la probabilidad con la que estas interacciones son requeridas de manera simultánea al observar distintos patrones de co-expresión (Han, et al., 2004) lo que no necesariamente significa que participan en el mismo proceso celular. En este sentido, Gomez y colaboradores intentaron identificar proteínas “moonlighting” a través de homólogos distantes en la base de datos del NCBI (<http://www.ncbi.nlm.nih.gov/protein/>) y en la base de datos no redundante de SWISS-PROT (<ftp://ftp.ebi.ac.uk/pub/databases/uniprot/knowledgebase/>) mediante algoritmos de alineamiento como PSI-BLAST (<http://blast.ncbi.nlm.nih.gov/>) y SAM (http://compbio.soe.ucsc.edu/SAM_T08/T08-query.html). Posteriormente analizaron tales regiones que presentan similitud con diferentes proteínas blanco mediante algoritmos específicos para identificar dominios o motivos relacionados con funciones particulares como PRODOM (<http://prodom.prabi.fr/>) y PFAM (<http://pfam.sanger.ac.uk/>). De tal forma, aunque ninguno de los programas resultó completamente acertado, se demostró su utilidad para identificar funciones secundarias que después pueden (y/o deben) ser analizadas mediante métodos bioquímicos (Gómez, et al., 2003). Adicionalmente, debido a que las proteínas intrínsecamente desordenadas tienen el potencial de modular su acción sobre diferentes moléculas (o incluso sobre la misma) dependiendo de la conformación que asumen tras su unión, el uso de los servicios IUPred (<http://iupred.enzim.hu/>) y PONDR VL-XT (<http://www.pondr.com/index>) demostró que el porcentaje de desorden estructural también puede ser relacionado con la multifuncionalidad de distintas proteínas (Tompa, et al., 2005).

Aquí se describe una aproximación alternativa a los enfoques bioinformáticos anteriores que puede ser útil para identificar proteínas involucradas en el metabolismo que presentan por lo menos dos actividades distintas y por lo menos una de ellas es indispensable para el crecimiento. De este modo, la identificación de proteínas multifuncionales en el metabolismo podría servir para explicar el papel indispensable de algunas enzimas metabólicas relacionadas con la proliferación celular en distintas condiciones. Por ello se revisó el papel de tres enzimas indispensables para el crecimiento en cuanto a los resultados de nuestro análisis y con respecto a la literatura disponible para analizar con más detalle algunas de estas predicciones. Estas enzimas fueron escogidas por tres criterios distintos: a) la

reductoisomerasa bifuncional de acetohidroxiácidos (ILV5) que fue reportada como multifuncional pero nuestro método no la detecta como indispensable, b) la fructosa 1,6 bifosfato aldolasa por haber sido reportada como "moonlighting" y predicha mediante nuestro enfoque tanto en la RIM como en la RIDM y, c) la triosa fosfato isomerasa como un caso nuevo de multifuncionalidad predicho por nuestra aproximación solo en la RIDM.

El GIC *YLR355C* produce la proteína "moonlighting" ILV5 que participa en la biosíntesis de los aminoácidos valina, leucina e isoleucina (Zelenaya-Troitskaya, et al., 1995). Sin embargo, se ha encontrado que la actividad enzimática correspondiente a esta proteína se reprime en la presencia de estos aminoácidos y presenta actividad residual incluso en las mutantes que no presentan este gen (Petersen y Holmberg, 1986). Este GIC es uno de los únicos tres genes de la **Tabla 9** implicados en la biosíntesis de aminoácidos (ILV1, ILV2 e ILV5) y se sabe que éstos también son reprimidos en medio rico. Por ello, una posible explicación para su indispensabilidad podría estar relacionada con su participación en la estabilidad del ADN mitocondrial (Zelenaya-Troitskaya, et al., 1995). Esto en relación a la complementación por sobreexpresión de ILV5 del fenotipo provocado por la eliminación del gen *ABF2* que genera células ρ^0 que crecen pobremente en medio rico. Sin embargo, como ya se había mencionado, la expresión del gen *ILV5* se reprime en medio rico y, consistente con ello, no fue clasificado como GIC en la RIDM sugiriendo que su multifuncionalidad dentro del metabolismo no es suficiente para explicar su papel indispensable.

El gen *YKL060C* codifica una enzima previamente identificada como "moonlighting" con actividad de aldolasa que cataliza la hidrólisis de fructosa 1,6 bifosfato en gliceraldehído-3-fosfato y dihidroxiacetona-fosfato. La expresión de mutantes con esta aldolasa inactiva han mostrado que la falta de esta actividad no permite el crecimiento cuando las levaduras crecen en glucosa pero esta condición se puede revertir cuando crecen en medio con etanol (Lobo, 1984). Alternativamente, se ha mostrado mediante interacciones físicas que esta aldolasa se requiere para el ensamblaje y la actividad de la ATPasa vacuolar (Lu, et al., 2001) y por eso se le considera como bifuncional. Si bien estos antecedentes muestran que esta aldolasa está relacionada con por lo menos dos funciones, ambas se reflejan en actividades enzimáticas: la hidrólisis de moléculas de fructosa 1,6 bifosfato y el consumo de ATP por la V-ATPasa. El hecho de que haya sido clasificado como GIC tanto en la RIM KEGG como en la RIDM Yipd.KEGG usando la centralidad SD muestra que la relaciones químico genéticas de esta aldolasa son suficientes para explicar su indispensabilidad incluso cuando se consideran las interacciones físicas que se establecen con los productos de los genes *VMA42/43/44*, *YMR123W*, *YKL119C*, *YGR105W*, *YHR060W*, *YGR106C*, *YPR036W*, *YMR054W*, *YOR270C*, *YCL005W-A*, *YDR328C*, *YBR127C*, *YKL080W*, *YEL051W*, *YOR332W*, *YGR020C*, *YHR039C-A*, *YLR447C*, *YDR202C*, *YJR033C* que se expresan como distintas subunidades y factores de ensamblaje de esta V-ATPasa. Además, se ha mostrado que cuando dichas subunidades son mutadas reducen el crecimiento vegetativo en presencia de glucosa (Giaever, et al., 2002; Winzeler, et al., 1999). De este modo aunque este GIC posee dos funciones y una de ellas depende de IPPs su clasificación no depende de la adición de interacciones físicas y su indispensabilidad tampoco parece depender de éstas.

También se clasificó el gen *YDR050C* como un GIC' que codifica para una proteína multifuncional. Este gen produce la proteína TPI1 con actividad de Triosa Fosfato Isomerasa y ya se ha discutido que la falta de su actividad enzimática no es letal en humanos (Ralsler, et al., 2006; Orosz, et al., 2009). Esto apoya la hipótesis con la que fue descrita en este trabajo en la que su actividad catalítica no explica su indispensabilidad. Además se ha reportado que presenta un fenotipo de letalidad sintética cuando este gen se elimina junto con el gen *YHR129C*, que codifica para la proteína ARP1 relacionada con la actina, lo que apoya aún más su multifuncionalidad en la levadura (Biogrid). En *Drosophila melanogaster* ya se desarrollaron sistemas de ingeniería genética que permiten la generación de de la proteína con distintas variantes de la actividad de TPI con las que se demostró *in vivo* que su deficiencia puede ser complementada con una TPI inactiva catalíticamente (Roland, et al., 2013). Estos resultados demuestran una función no metabólica para la TPI cuya pérdida contribuye significativamente a la disfunción neurológica en un modelo de mosca planteando como hipótesis que la presencia de la enzima es crítica, independientemente de la actividad catalítica. Esta identificación concuerda con el hecho de que solo pudo ser clasificado como GIC al añadir IPPs de la base de datos YIPD sobre las interacciones metabólicas red KEGG y abre nuevas vías de investigación para probar el rol de la TPI en el crecimiento.

De manera importante, la aldolasa bifuncional codificada por el GIC *YKL060C* ha sido descrita como una enzima esclava (Teusink y Westerhoff, 2000) que no puede ejercer ningún tipo de control sobre el flujo metabólico de la glucólisis y la TPI codificada por el GIC *YDR050C* juega un papel relacionado a tal característica. En breve, las llamadas enzimas esclavas no controlan más que sus correspondientes metabolitos esclavos al participar en vías lineares en donde una enzima no es sensible a su producto por lo que la siguiente enzima es la única sensible a éste. Lo anterior cobra importancia cuando las enzimas río abajo de la enzima esclava forman una unidad monofuncional (Rohwer, et al., 1996) en la que 1) los metabolitos que quedan dentro de tal unidad no pueden regular enzimas que están fuera de tal unidad, 2) existe sólo una forma de conectar el flujo hacia el modulo con el resto del metabolismo y 3) no existe conservación parcial de materia (como en el caso de los nucleótidos de Adenina en donde sólo se transfiere fosfato). Así, debido a que la fosfofructocinasa es casi completamente insensible a su producto, la fructosa 1,6 bifosfato (Otto, et al., 1986), la aldolasa que es la enzima siguiente en la glucólisis se puede considerar como esclava por lo que tiene poco o ningún control sobre el flujo glucolítico. Al analizar bajo esta perspectiva las demás enzimas que siguen a este paso en la glucólisis (siendo la primera de ellas la TPI) es importante notar que la bifurcación que sigue a este paso impide agrupar las enzimas subsecuentes en unidades monofuncionales, sobre todo si el flujo hacia la producción de glicerol, trehalosa o glucógeno es significativo. Así, en el caso de la rama que va hacia la producción de glicerol, si presenta un flujo significativo sólo la aldolasa pierde la capacidad de controlar cualquier otra variable glucolítica, a excepción de la concentración de fructosa 1,6 bifosfato. Por consiguiente, las enzimas subsecuentes deberían poder ejercer control sobre el flujo glucolítico pero las únicas variables que podrían controlar serían las concentraciones de los nucleótidos de Adenina (e.g. NADH) de manera indirecta disminuyendo las ya de por sí limitadas capacidades de control.

Lo anterior significa que las proteínas codificadas tanto por el GIC *YKL060C* como por el GIC *YDR050C* no serían capaces de controlar directamente el flujo glucolítico limitando así la importancia de su actividad catalítica y resaltando el papel de cualquier otra función que pudieran desplegar cuando la célula crece en medio con glucosa. Así, por un lado se podría explicar la importancia de la fosfofructocinasa en la regulación del flujo glucolítico en relación a la existencia de por lo menos 7 distintas proteínas asociadas a tal función. Por otro lado, tanto la descripción de las modificaciones del flujo glucolítico ante la eliminación de estos GICs *YKL060C* y *YDR050C* como la descripción de las funciones de las proteínas codificadas por estos genes podría revelar nuevas relaciones regulatorias de la unidad metabólica en la que participan. De tal manera, la eliminación del GIC *YKL060C* elimina la estructura de la vía metabólica que relega todo el control del flujo glucolítico hacia la fosfofructocinasa resaltando la importancia de su función catalítica. En contraste, la eliminación de la TPI no modifica tal bifurcación y resalta la importancia de su función adicional que podría incluir las distintas IPPs en las que participa u otro tipo de relaciones funcionales.

IV.d ANÁLISIS *IN VIVO*

La noción de que las enzimas pueden llevar acabo múltiples funciones que pueden resultar indispensables para el crecimiento como resultado de sus interacciones físicas además de su actividad catalítica ya fue discutida en las secciones anteriores. Sin embargo, por lo menos en el caso de las enzimas metabólicas codificadas por GICs, aún no se ha descrito si su indispensabilidad está relacionada directamente con su actividad catalítica o si puede ser el resultado de otro tipo de interacciones. Los GICs' clasificados tras la integración de IPPs sobre las RIMs dan indicios de la importancia de las uniones físicas como función adicional a la actividad catalítica y presentan una oportunidad interesante para preguntar si es que realmente presentan dos o más funciones y cuál de ellas determina el fenotipo indispensable para el crecimiento.

Para confirmar distintas funciones en las proteínas generalmente se requieren enfoques experimentales complementarios, sobre todo si tales conclusiones fueron obtenidas mediante tecnologías de análisis masivo o si fueron inferidas por métodos computacionales. Cuando la meta de validación incluye confirmar conclusiones biológicas específicas se puede optar por utilizar las características con mayor interés biológico. Sin embargo, cuando se espera validar la significancia del método o de las listas

de resultados no existe una aproximación estándar de validación por lo que es posible: 1) validar solo los mejores resultados basados en la importancia biológica o estadística, 2) validar todos los resultados o, 3) validar solo un conjunto pequeño aleatorio (Leek, et al., 2012). En este sentido, los GIC's aislados a partir de la red Yipd.KEGG2 con la centralidad SD presentan tanto actividad catalítica como contactos físicos con otras proteínas. Una alternativa para saber si la indispensabilidad de estos genes depende de su actividad catalítica, de otro tipo de interacciones funcionales o de ambos factores consiste en eliminar una sin alterar la otra. En el caso de las IPPs no siempre se conocen todos los interactores de una proteína determinada, ni mucho menos los aminoácidos necesarios para dicha interacción. En contraste, los sitios catalíticos de un gran número de enzimas ya han sido descritos extensivamente tanto por técnicas bioquímicas como de biología molecular utilizando inhibidores específicos o mutagénesis (respectivamente). Sin embargo, aunque se han sintetizado un buen número de inhibidores, éstos no han sido generados para todos y cada uno de los sitios catalíticos descritos. Por otro lado, los que se encuentran disponibles suelen ser estudiados *in vitro* con las proteínas purificadas limitando su uso *in vivo*. Por tanto, la mutagénesis dirigida sobre sitios catalíticos plantea una mejor alternativa para abolir la actividad enzimática de dichos genes. Más aún, dada la alta estandarización de los métodos de biología molecular y de análisis bioquímico además de la diversidad de colecciones de herramientas desarrolladas para *S. cerevisiae* existe una alta probabilidad de que los protocolos utilizados y los resultados obtenidos con estos pudieran ser fácilmente extrapolables a los demás GICs.

En este caso se eligió el GIC *YDR050C* para analizar la función biológica de este tipo de genes y validar la utilidad del método aquí presentado mediante los procedimientos descritos en el apartado de **Estrategias *in silico*** y los antecedentes mencionados en el apartado de **Estrategias *in vivo*** que también se discuten en el último párrafo de la sección anterior. Así, para evaluar el fenotipo provocado por una proteína sin actividad catalítica producida a partir del *YDR050C* no fue posible partir de una cepa haploide nula por tratarse de un GIC. En este sentido, una de las ventajas de *S. cerevisiae* como modelo de estudio consiste en que su ciclo vital incluye etapas en las que se puede aislar tanto en forma haploide como diploide y existen dos cepas diploides heterocigotas (Y23986 y Y26690; **Tabla 4**) provenientes de la cepa BY4743 con el genotipo MAT *a/α*; *YDR050C::kanMX4/ YDR050C*. Estas mutantes si pueden crecer en medio rico a diferencia de la haploide nula para el gen *YDR050C* ya que presentan la delección de un GIC (cuya secuencia fue intercambiada por un cassette KanMX4 de resistencia a genética mediante la misma estrategia de delección basada en PCR y recombinación utilizada por el SGDP) pero también presentan una copia silvestre del mismo gen.

Tales cepas fueron sometidas a varios procedimientos descritos en la sección **II.h** para obtener una cepa nula para el gen *YDR050C* en la cual se podría analizar el fenotipo causado por la mutación presente en el plásmido MoBY_TPI1_E97Q y su efecto en el crecimiento. Sin embargo, aun habiendo recuperado distintas cepas en los medios selectivos, al confirmar su genotipo mediante PCR de colonia, se encontró que la selección de haploides con canavanina resultó efectiva en la mayoría de los casos pero ninguna de las haploides seleccionadas presentaron solo el cassette KanMX4. Lo anterior debido a varios factores como el hecho de que uno de los métodos utilizados dependía de la selección de esporas que es difícil y poco eficiente. Este proceso pudo haberse optimizado al utilizar selección manual con micromanipuladores. Sin embargo, esta no sería una solución completa ya que aún requeriría de la selección de las mutantes de interés. En este punto se encontró otra limitante no prevista debido a que el vector con el gen mutante utiliza la resistencia a G418 como marcador de selección y es el mismo con el que se deberían seleccionar las haploides nulas para el gen *YDR050C* obtenidas a partir de las cepas Y23986 y Y26690. Lo anterior dificultó la selección de células haploides nulas para el gen *YDR050C* previamente transformadas con los plásmidos MoBY_TPI1_wt o MoBY_TPI1_E97Q aún cuando fueron sometidos a selección en medio sin uracilo (**Tabla 6**) ya que contienen el marcador URA3 de prototrofia por uracilo como se muestra en la **Figura 1**. Esto puso en evidencia nuevamente los problemas que ya se mencionaron con el método de selección de esporas, ya sea porque no se obtuvieron células haploides o porque su selección con canavanina no fue eficaz.

Debido a esto se probó una estrategia adicional utilizando el método de eliminación basada en PCR descrito en la sección **II.i** sobre las cepas BY4741y BY4742 (**Tabla 4**) que intentaron recuperarse en los distintos medios utilizados para recuperar esporas (**Tabla 6**). Lo anterior debido a que aunque se ha descrito que la remoción de un GIC como *YDR050C* impide el crecimiento en medio con glucosa (YPD)

de células derivadas de la cepa S288C tanto en las mencionadas formas haploides (BY4741, BY4742) como en la doble eliminación en la cepa diploide correspondiente (BY4743) (Giaever, et al., 2002). Sin embargo, existen estudios en los que se ha mostrado que el crecimiento de mutantes de este tipo se puede recuperar en presencia de inositol para las cepas SMY10 y SMY15 (Shi, et al., 2005), en medio YPEG para la cepa W303 (Compagno, et al., 1996) y en medio YPED para la cepa CEN.PK 113-7D (CBS8340) (Compagno, et al., 2001). Entonces, se generaron fragmentos de PCR que presentan extremos homólogos al gen *YDR050C* pero incluyen el "cassette" KanMX4 en el lugar del ORF. Dichos fragmentos se utilizaron por separado para transformar las cepas BY4741(MAT a) y BY4742(MAT α) mediante el método de (Gietz y Woods, 2002) para después recuperar las células transformantes en los medios de cultivo YPEG, YPED y SC+ino descritos en la **Tabla 6** de la sección II.i. Con este último procedimiento se recuperaron varias colonias en medio SC con inositol y G418 verificando falta de crecimiento en el medio de selección donde fueron sembradas las cepas sometidas al proceso de transformación sin fragmento de ADN. Sin embargo, al utilizar el medio SC con sulfato de amonio y G418 se observó crecimiento incluso al sembrar las cepas sometidas al proceso de transformación sin ADN exógeno. Esto tiene que ver con la modificación del medio indicada en la sección de métodos en la que fue sustituido el sulfato de amonio por el glutamato monosódico debido a que el primero interfiere con la acción de este antibiótico. Una vez recuperadas las colonias presentadas en el panel superior derecho de la **Figura 9** fueron utilizadas para iniciar nuevos cultivos ya sea en medio sólido o líquido pero no presentaron crecimiento y también se utilizaron para hacer PCRs de colonia pero tampoco presentaron el genotipo esperado mostrando que esta alternativa tampoco resultó efectiva.

Puesto que se esperaba que el análisis de esta mutante fuera difícil por tratarse de un GIC cuya eliminación impide el crecimiento en condiciones estándar se trató de recuperar en los medios en los que se había reportado crecimiento para células deficientes en este gen. Sin embargo, en el caso de la complementación del crecimiento por inositol las células mutantes de tales estudios no se obtuvieron mediante el mismo método de transformación y eliminación sino que se aislaron de una biblioteca de células sometidas a mutagénesis química con etilmetano sulfonato (Lindgren, et al., 1965). En el caso de las células recuperadas en los medios YPE y YP se utilizaron distintas cepas (W303 y CEN.PK 113-7D (CBS8340)) que se transformaron mediante distintas variaciones del método de disrupción en un paso (Rothstein, 1991; Klebe, et al., 1983; Ito, et al., 1983). La estandarización de estos distintos métodos de mutagénesis, transformación y/o recuperación de células van más allá de los objetivos de esta tesis y no fueron explorados. Sin embargo, es posible proponer el uso de otro método en el que se utiliza la recombinasa Cre y sitios lox para reemplazar un gen de interés con oligos que permiten amplificar por PCR tanto las secuencias a los extremos del gen de interés como un "cassette" de resistencia y los sitios loxP a los extremos (Gueldener, et al., 2002). La ventaja de este método radica en que, al reemplazar el gen de interés, es posible activar la expresión de la recombinasa Cre eliminando posteriormente los sitios loxP y la secuencia que queda entre ellos además del "cassette" de resistencia. Esto permitiría utilizar el mismo método de selección por G418 de manera secuencial reutilizando la resistencia y evitando que la presencia de un "cassette" interfiera con el fenotipo provocado por el mismo "cassette" en otro *loci* (Hegemann y Heick, 2011). Además, este protocolo permitiría lidiar con el problema antes mencionado en el que tanto las cepas diploides heterocigotas Y23986 y Y26690 como las que contienen el plásmido MoBY_TPI1_E97Q utilizan el mismo método de selección.

Entonces, aunque esta proteína mutante no se pudo expresar en el fondo genético adecuado se logró obtener un vector en el que encuentra un GIC mutante que debería expresar una proteína sin actividad de TPI. Así, aún cuando este análisis no arrojó resultados concluyentes, mostró ser potencialmente útil para estudios posteriores. Por ejemplo, ya se mencionó que la posible multifuncionalidad de las enzimas críticas para el metabolismo se puede estudiar mediante experimentos de mutagénesis y complementación de fenotipos. Otro enfoque podría resultar del estudio de interacciones genéticas en donde se infiere la función de un par de genes en base al fenotipo que resulta tras la remoción de ambos (Tong, et al., 2001) que incluso podrían ser acoplados a microarreglos (Ooi, et al., 2003). Una vez estandarizado, este método puede ser aplicado para conjuntos de genes seleccionados en base a algún tipo de correlación funcional (Schuldiner, et al., 2005). Lo que es común a estas técnicas es que una parte significativa de los casos en los que se encuentran interacciones entre un par de genes, éstos participan en complejos multiprotéicos evaluando el fenotipo obtenido con la cruce de mutantes con doble eliminación en células homocigotas. La mutante sin actividad catalítica

generada en este estudio podría ser utilizada para investigar cómo se modifican las interacciones genéticas del complejo en el que se sabe que participan (tomando sus vecinos y las relaciones entre ellos dentro de la red REMG) para evaluar más a fondo la participación de las IPP para determinar el fenotipo crítico.

Por último el gen *YDR050C* se ha implicado en la producción de glicerol mediante varios estudios (Overkamp et al. 2002; Cordiere et al. 2007). Dicho compuesto es de importancia industrial y se ha demostrado que en cepas que tiene eliminado este gen se puede incrementar su producción. Sin embargo, al tratarse de un gen crítico para el crecimiento en glucosa, aún cuando se pueda aumentar la producción de este metabolito el crecimiento del organismo presenta deficiencias que dificultan su explotación industrial. Si la mutante construida en este estudio logra restablecer el crecimiento del organismo en YPD podría ser una opción para distintas aplicaciones biotecnológicas.

IV.e OTRAS LIMITACIONES Y ALTERNATIVAS

Esta unión de redes que relacionan genes usando diferentes tipos de actividades podría haber resultado en la adición de una interacción del tipo proteína-proteína completando lo que debería ser del tipo químico-genética. En este sentido, 15 de las interacciones de la red Yipd también se encuentran en la red KEGG2path y 14 de ellas pertenecen a la misma vía además de que comparten metabolitos distintos a los que fueron removidos como se indica en la **Tabla 1** aunque ninguna de estas interacciones involucra algún GIC. Adicionalmente, al tomar las 8756 interacciones de la red Yipd que no están representadas en la red KEGG2path solo 64 (0.73%) presentan metabolitos en común (siguiendo las mismas restricciones de metabolitos eliminados y vías en común que se indican en la **Tabla 1**) entre las cuáles nada más seis incluyen por lo menos un GIC. La información anterior se obtuvo utilizando el API de la base de datos KEGG (<http://www.kegg.jp/kegg/rest/keggapi.html>). Es importante notar que para interpretarla es necesario tomar en cuenta que esta base de datos se encuentra en constante actualización por lo que las versiones más recientes incluyen unas cuantas interacciones que no se consideraron anteriormente. Además, algunas interacciones de las RIDs utilizadas para enriquecer las distintas RIMs no necesariamente están relacionadas con en el metabolismo ni con los GICs implicados en éste y, por ello, existe la posibilidad de que la mejoría en la clasificación fuera resultado del sobreajuste provocado por el exceso de interacciones sin relevancia aparente. Sin embargo, al utilizar el conjunto Unión de las RIDs se obtuvieron valores más bajos en comparación con el uso de las distintas RIDs que lo componen pero de manera individual. Esto sugiere que las observaciones anteriores no son un artefacto ya que, aunque tiene más interacciones, no mejoran tanto la clasificación de GICs como las redes construidas con interacciones individuales. Así, se puede notar que las interacciones incluidas pueden modificar la clasificación de GICs en las distintas redes y sugieren nuevamente la necesidad de actualizar y/o depurar de las interacciones incluidas dentro de la red.

Una aproximación para buscar la depuración de genes y/o interacciones de la red REMG tiene que ver con la naturaleza de la medida con la que se obtuvieron los mejores resultados. De tal forma, si se toman ya sean todos los GICs o solo los clasificados usando la centralidad SD para seleccionar sus vecinos a uno y dos nodos de distancia dentro de la red Yipd.KEGG2path deberían obtenerse resultados por lo menos semejantes a los que se obtienen con la red completa. El procedimiento anterior resulta en un $ABC=0.754\pm 0.024$ usando todos los GICs y un $ABC=0.798\pm 0.021$ usando solo los GICs correctamente clasificados (casos 1 y 2 en la **sección A** de la **Tabla 10**). Así, los valores de ABC obtenidos usando todos los GICs no son significativamente mayores a los conseguidos con el par Yipd.KEGG2path-SD ($ABC=0.734\pm 0.025$ ver **Tabla 7** y la Tabla C3 complementaria) e incluso en el segundo caso son significativamente mayores. Sin embargo, aún cuando los valores de ABC son iguales o mayores al considerar solo los GICs en cuestión se producen redes que son muy distintas a la original. En este sentido, la red Yipd.KEGG2path tienen 3570 vértices y 16886 aristas mientras que las redes construidas a partir de las interacciones locales de los GICs totales o clasificados correctamente cuentan con 1854 y 1835 vértices en 5639 y 5596 aristas respectivamente representando ~50% de los vértices y ~33% de las aristas de la red original. Además, estas redes son muy distintas debido a que solo en la que se incluyeron todos los GICs se pueden encontrar dos componentes conectados (subredes en las que

existe un camino entre cada uno de los genes que las componen) mientras que al incluir solamente los GICs predichos no existe ninguno. Este efecto diferencial señala que los GICs que no pueden clasificarse correctamente con el par Yipd.KEGG2path podrían jugar un papel reflejado en la centralidad global de la red a diferencia del papel local de los GICs que si fueron correctamente clasificados.

Tabla 10

etiqueta	ABC	99%ls	99%li	error	sensibilidad	especificidad	exactitud
Yipd.KEGG2path-SD	0.734	0.759	0.709	0.465	0.813	0.574	0.962
A							
1:	0.754	0.778	0.730	0.402	0.963	0.599	0.928
2:	0.798	0.819	0.777	0.367	0.872	0.656	0.941
B							
Biogrid	0.606	0.640	0.572	0.584	0.649	0.534	0.978
Dip	0.760	0.788	0.733	0.444	0.724	0.653	0.968
Intact	0.659	0.696	0.623	0.550	0.604	0.618	0.977
Mpact	0.742	0.766	0.718	0.449	0.754	0.625	0.963
Union	0.677	0.708	0.647	0.522	0.731	0.552	0.979
Yeastnet	0.803	0.829	0.776	0.395	0.791	0.665	0.975
Ypi	0.736	0.761	0.711	0.460	0.724	0.632	0.963

Depuración y adición de interacciones. Se muestran los valores de los parámetros utilizados para evaluar el desempeño de la centralidad SD en la clasificación de los conjuntos de GICs en las redes que se indican a continuación. En la primer fila se muestran los valores obtenidos usando los GICs de la red KEGG2path y la red Yipd.KEGG2path como referencia. Para obtener los valores de la sección **A** se utilizó la subred de vecinos a uno y dos nodos de distancia de 1: todos los GICs en KEGG2path y 2: solo aquellos GICs que son correctamente clasificados usando el par Yipd.KEGG2path-SD. En la sección **B** se muestran los valores obtenidos con el conjunto unión de la red Yipd.KEGG2path y la subred de vecinos a 1 y 2 nodos de distancia de los falsos positivos obtenidos con el par Yipd.KEGG2path-SD seleccionados a partir de la red indicada en la fila correspondiente. Los casos en los que se mejoró el desempeño con respecto a los valores de referencia se indican en negritas.

Aún así, es posible usar una vez más las bases de datos de **Tabla 2** para seleccionar las relaciones a uno y dos nodos de distancia de todos los GICs no clasificados y añadirlas a la red Yipd.KEGG2path para clasificar los GICs que no fueron predichos previamente con la centralidad SD. De manera interesante, solo en el caso de las interacciones obtenidas a partir de Yeastnet se pudo obtener un valor de ABC significativamente mejor ($ABC=0.803\pm 0.026$; **Tabla 10 sección B**) pero aún lejos de clasificar correctamente todos los GICs. Este último dato señala que la adición de algunas interacciones genéticas probabilísticas contenidas en Yeastnet pueden mejorar el desempeño del par Yipd.KEGG2path-SD aunque no queda claro cuál sería la implicación funcional de este fenómeno. Es por ello que estos 25 GICs que no fueron clasificados con el par Yipd.KEGG2path-SD (ver **Tabla 8 y Figura 11**) se analizaron un poco más a detalle y se encontró que están relacionados con reacciones que participan en alguna de las siguientes vías: [sce01110](#) Biosynthesis of secondary metabolites (14), [sce00900](#) Terpenoid backbone biosynthesis (6), [sce00860](#) Porphyrin and chlorophyll metabolism (3), [sce00562](#) Inositol phosphate metabolism (3), [sce04070](#) Phosphatidylinositol signaling system (3), [sce00520](#) Amino sugar and nucleotide sugar metabolism (2), [sce00100](#) Steroid biosynthesis (2), [sce00970](#) Aminoacyl-tRNA biosynthesis (2), [sce04144](#) Endocytosis (1), [sce00770](#) Pantothenate and CoA biosynthesis (1), [sce01210](#) 2-Oxocarboxylic acid metabolism (1), [sce04113](#) Meiosis (1), [sce04146](#) Peroxisome (1), [sce00909](#) Sesquiterpenoid and triterpenoid biosynthesis (1), [sce00290](#) Valine, leucine and isoleucine biosynthesis (1), [sce01230](#) Biosynthesis of amino acids (1), [sce00600](#) Sphingolipid metabolism (1), [sce00740](#) Riboflavin metabolism (1), [sce00790](#) Folate biosynthesis (1). Más allá, se encontró que entre estos genes se encuentran sobrerrepresentados 12 términos GO (The Gene Ontology Consortium, 2004) para proceso biológico obtenidos al contrastarlos contra la anotación funcional del resto del genoma de *S. cerevisiae* utilizando la herramienta FatiGO (Al-Shahrour, et al., 2004) del servidor BABELOMICS (Medina, et al., 2010) con un p-value de 0.005. Estos términos concuerdan con las vías mencionadas anteriormente y son isoprenoid biosynthetic process ([GO:0008299](#); adj. p-value= 2.501×10^{-9}), lipid metabolic process ([GO:0006629](#); adj. p-value= 1.037×10^{-8}), lipid biosynthetic process ([GO:0008610](#); adj. p-value= 1.078×10^{-8}), steroid biosynthetic process ([GO:0006694](#); adj. p-value= 1.251×10^{-8}), steroid metabolic process ([GO:0008202](#); adj. p-value= 5.801×10^{-8}), sterol biosynthetic process ([GO:0016126](#); adj. p-value= 1.468×10^{-7}), alcohol metabolic process ([GO:0006066](#); adj. p-value= 1.898×10^{-7}), sterol metabolic process ([GO:0016125](#); adj. p-value= 4.382×10^{-7}), cellular nitrogen

compound biosynthetic process ([GO:0044271](#); adj. p-value=0.000747), heme biosynthetic process ([GO:0006783](#); adj. p-value=0.0007692), porphyrin-containing compound biosynthetic process ([GO:0006779](#) adj. p-value=0.001806), tetrapyrrole biosynthetic process ([GO:0033014](#); adj. p-value=0.001806).

Al analizar estas vías se puede observar que una parte importante de los GICs no clasificados participan en vías relacionadas con la biosíntesis de porfirinas lo que es congruente con su relación ontológica con el metabolismo de alcohol, procesos biosintéticos de grupos hemo, de compuestos que contienen porfirinas y de tetrapirroles. En otro grupo se encuentran vías de biosíntesis de esteroides, de pantotenato y Co-A además de esfingolípidos que son relacionadas con términos GO de biosíntesis y metabolismo de isoprenoides, de lípidos, de esteroides y esterolés. En un último grupo se incluyen vías de metabolismo de aminoácidos (en particular valina, leucina e isoleucina entre los que se encuentra el GIC multifuncional no clasificado *YLR355C/ILV5* discutido en la sección **IV.c**) congruentes con términos GO de procesos celulares de biosíntesis de compuestos celulares con nitrógeno. Estas relaciones indican que además del papel estructural en la conectividad global que ya se había discutido tienen un correlato funcional con el metabolismo aeróbico por lo que no necesariamente deberían estar relacionadas con el crecimiento en medio con glucosa invitando a cuestionar su indispensabilidad más a fondo.

Otra alternativa para explicar las limitaciones en la clasificación GICs metabólicos podría incluir el considerar que la función de éstos es consecuencia de un proceso dinámico y no estructural como lo simulan las medidas de centralidad. En este sentido, los modelos cinéticos que implementan ecuaciones diferenciales son parte de métodos bien establecidos para analizar vías bioquímicas (Segel, 1975). Así, se podría suponer que una medida que evalúa solo la topología de la red no es capaz de distinguir correctamente los GICs debido a que su indispensabilidad depende de la velocidad de reacción de las enzimas que codifican. También se han modelado distintos procesos celulares de manera efectiva usando modelos dinámicos en los que se reemplaza el uso de ecuaciones diferenciales por relaciones estocásticas para describir reacciones químicas (Gillespie, 1976). Estos modelos se basan en reglas y restricciones involucradas con la conservación de la materia y han sido refinados hasta el punto en el que consideran la estequiometría de las reacciones (Análisis de Control Metabólico) y/o el flujo a través de las distintas reacciones (Análisis de Balance de Flujos) que se ha propuesto que podrían llevar a un entendimiento sistémico del metabolismo, aún cuando presentan limitaciones relacionadas al número de variables y la necesidad de estimación de parámetros (Klipp, 2007). Aún así, existen tendencias tanto por simplificar las redes en base a módulos y/o complejos (Kaltenbach y Stelling, 2012) como por extrapolar los métodos a modelos más grandes (Kent, et al., 2012). De este modo, las redes reconstruidas en éste trabajo podrían ser utilizadas para formular modelos a escala genómica y analizarlos con métodos optimizados o bien, los complejos aquí descritos en base a los vecinos de los GICs y las interacciones entre ellos podrían resultar en una forma interesante de simplificar redes metabólicas para analizarlas en base a su estequiometría y tomando en cuenta la modularidad estructural que aquí se presenta.

V. CONSIDERACIONES FINALES

La operación de la célula es orquestada mediante acciones regulatorias y de retroalimentación que ocurren de manera jerárquica sobre componentes celulares pertenecientes a distintas escalas de magnitud y tiempo. Lo anterior resulta en procesos organizados en distintos niveles cuya función es tan compleja que, aún cuando existe una gran cantidad de información acerca de ellos, todavía se desconoce la totalidad de mecanismos que los controlan. Tal es el caso del metabolismo que, aunque ha sido sugerido en diversas ocasiones como un buen punto de partida para las tareas de la biología de sistemas, podría resultar siendo el proceso celular más difícil de entender a nivel de sistema (Heinemann y Sauer, 2010).

Pese a lo anterior, se ha propuesto que las técnicas de experimentación a gran escala o de alto rendimiento y los algoritmos computacionales utilizados para analizar los datos arrojados mediante dichas técnicas podrían llevar a la descripción sistémica del metabolismo. Es por ello que cada vez hay más estudios, como éste, en los que se integra información de distintos tipos considerando procesos biológicos como sistemas complejos en los que hay varias partes que pueden interactuar de tal manera que el todo presenta características que no presentan sus partes aisladas. Así, podría considerarse el 'íntegroma' con información de todas las '-ómicas' además de otros tipos de información relevante en un mismo contenedor para realizar análisis integrados buscando los mejores resultados pero pareciera que para develar los misterios más grandes de la biología, antes que inventar nuevas '-ómicas', es necesario combinar las que ya existen (Baker, 2013).

Esto forma parte de una tendencia por rellenar espacios vacíos al encontrar patrones similares de organización para describir genes con funciones particulares. Incluso, cuando la junta editorial de BMC Biology invitó a sus miembros a describir preguntas importantes o interesantes para la Biología para celebrar su decimo aniversario, Stephen Benkovic planteó la pregunta de ¿cómo encontrar el significado funcional de interacciones proteicas a escala genómica?. Esto debido a la reciente proliferación de publicaciones que describen la identificación de complejos citosólicos multiprotéicos en varios organismos que pueden actuar como módulos funcionales con distintas actividades biológicas. Adicionalmente, Dagmar Ringe planteó ¿cómo se puede ir de moléculas a organismos? esto debido a que los estudios evolutivos están enfocándose hacia la predicción de la organización de organismos, como se relacionan sus distintas partes entre si y como el nivel molecular se transforma en nivel de organismo (Benkovic, et al., 2013).

Tomando en cuenta lo anterior, los estudios como el que aquí se presenta han abordado la posibilidad de ligar relaciones a escala molecular con comportamientos a escala de organismos considerando nociones de complejidad y esperando entender como emergen distintos comportamientos cuando sus distintos componentes interactúan. Así, se entiende que para obtener un panorama completo se deben integrar y combinar distintos estudios combinando enfoques teóricos para abrir la posibilidad de reconstruir los orígenes evolutivos de la organización celular. Sin embargo, todavía no existe un marco conceptual unificado para abordar las preguntas sobre la complejidad y tampoco se sabe en específico el tipo de información, cuanta se requiere o cuáles son las preguntas críticas a preguntar. Aún así, si bien nunca se alcanzaran a describir los sistemas complejos a detalle, las descripciones particulares que lleven a predicciones de características indispensables siguen estando al alcance (West, 2013). Tal es el caso de esta tesis en la que, aunque no se describe el funcionamiento del metabolismo celular en su totalidad, se describe como identificar GICs abriendo la posibilidad para describirlos en organismos y/o ambientes diferentes para utilizarlos con distintos propósitos.

Como comentario final, se ha encontrado que existen proteínas que son indispensables solo para ciertas condiciones ambientales y son perdidas cuando éstas cambian mientras que otras proteínas son indispensables para procesos básicos de la vida y son protegidas efectivamente contra la delección (Pawlowski, et al., 2013). Así, aún cuando en esta tesis solo se considera la condición estándar de crecimiento, 31 de estos GICs fueron estudiados en 14 diferentes condiciones experimentales y 30 fueron reportados como indispensables en por lo menos una de ellas sugiriendo que también pudieran ser requeridos en otras condiciones de crecimiento (Mnaimneh, et al., 2004).

Lo anterior podría ser llevado más allá considerando la extensa variedad de levaduras dentro del género *Saccharomyces*, e incluso de distintas cepas de la misma especie *cerevisiae* de uso industrial, cuyas secuencias genómicas ya han sido determinadas abriendo el camino para ensamblar y analizar su 'pan-genoma' (*i.e.* el conjunto completo de genes para una especie particular) revelando conjuntos de genes cuyas características pueden ser comunes o distintivas entre especies (Dunn, et al., 2012). Así, los conjuntos de GICs podrían ser utilizados con distintos fines dependiendo de las condiciones de crecimiento y/o las cepas en las que se presenten de manera que podrían ser utilizados, por ejemplo, para revelar la existencia de algún tipo de jerarquía de esencialidad dependiendo de la variedad de condiciones de crecimiento en los que se encuentran (*e.g.* los GICs que son indispensables en mayor número de condiciones ocupan un nivel más alto en la jerarquía que aquellos que se requieren en menos condiciones). Mientras tanto, aquellos GICs específicos para alguna cepa y/o condición de crecimiento podrían revelar especializaciones que pudieran ser explotadas con fines industriales, atacadas como blancos farmacológicos o utilizadas como puntos de control. Tales similitudes y diferencias entre cepas de la misma especie o entre diferentes géneros también podrían ser útiles para estudiar la evolución de estos organismos tomando la conservación o pérdida de GICs como punto de referencia.

VI. CONCLUSIONES

Al realizar este procedimiento se presentaron 1311 pares RIDM-centralidad con $ABC > 0.5$ en contraste con las 14 de 198 posibilidades (7.07% del total) en las que se presentó un $ABC > 0.5$ usando RIMs. Así, las RIDMs son hasta un 30% mejores que las RIMs obteniendo valores de ABC de hasta 0.9. Esto resalta la importancia de incluir las interacciones de las RIDs para completar las RIMs mejorando la clasificación de GICs. Con esto se corroboró que las RIDMs evaluadas según centralidades individuales son más útiles para distinguir GICs que cualquiera de las RIMs o RIDs originales por separado. Estos resultados apoyan el supuesto de que al considerar distintos tipos de interacciones es posible obtener una representación del metabolismo que permite la clasificación de GICs a partir de medidas estructurales.

En el 14% de los pares RIDM-centralidad la mejoría en la predicción de GICs fue debida al incremento en el valor de centralidad por la unión de RIMs y RIDs sin perder la capacidad predictiva que ya se tenía en la RIM. Con esto se pudo notar que las mejores RIDMs fueron construidas a partir de cualquier RIM unida con alguna de las RIDs Intact, M_{pact}, Y_{pd}, Y_{pi} y Union todas ellas presentando IPPs. Además, a excepción de 12 casos en los que fue utilizada la centralidad Ecclnv y 3 en los que se utilizó Trav, en este grupo aparecieron todas las centralidades locales. Esto es congruente con el tipo de interacciones añadidas y relaciona la esencialidad de los genes metabólicos con la existencia de funciones enzimáticas asociadas en complejos mediante interacciones físicas que brindan cierta noción de modularidad.

Los resultados de esta tesis apoyan el supuesto de que al integrar distintos tipos de interacciones en una misma red es posible obtener una representación del metabolismo que permite la clasificación de GICs a partir de medidas estructurales. Por ello, se elaboró un índice CIC para identificar aquellos GICs contenidos en al menos una de las RIMs que fueron clasificados exclusivamente al añadir interacciones y aumentar su centralidad en la RIDM correspondiente. Así se encontraron 56 GICs' que fueron predichos al añadir interacciones que además aumentaron su centralidad en la RIDM Y_{pd}.KEGG2 y seleccionados tomando en cuenta la centralidad SD obteniendo un $ABC = 0.734 \pm 0.025$ y un $CIC = 1$. Lo anterior asocia estos 56 GICs' con por lo menos dos funciones distintas al clasificarlos por aumento de centralidad debido a interacciones de RIDs sobre RIMs; con este par se clasifica el mayor número de GICs con el menor número de filtros a partir de la información de la base de datos KEGG (109 de 134 totales).

Se encontró que algunos de estos GICs' asociados a más de una función presentan un solo dominio representando más de la mitad de los GICs' predichos y planteando la existencia de GICs que se expresan como proteínas multifuncionales con un solo dominio. Además, como una alternativa complementaria para realizar una búsqueda de proteínas multifuncionales, se buscaron genes que producen proteínas tipo "moonlighting" mediante técnicas de minería de textos pero solo se pudieron capturar los GICs *YKL060C* y *YLR355C*. Tanto los resultados anteriores del análisis de dominios PFAM como los obtenidos por minería de textos muestran que no es trivial relacionar un gen con proteínas que llevan a cabo distintas funciones.

Mediante un ejercicio sistemático se identificaron nuevas propiedades de las redes metabólicas que no se habrían inferido mediante los enfoques tradicionales en los que se considera la función de un gene en un contexto aislado. Dichas propiedades incluyen la posible multifuncionalidad de enzimas metabólicas que también llevan a cabo IPP lo que hasta ahora deja abierta la pregunta de cuál de ellas es la que determina el fenotipo crítico.

Se eligió el GIC *YDR050C* para analizar la función biológica de este tipo de genes y se logró obtener un vector en el que encuentra una versión mutante de este GIC que debería expresar una proteína sin actividad de TPI. Aunque esta proteína mutante no se pudo expresar en el fondo genético adecuado se detectaron los problemas relacionados con distintos métodos de mutagénesis, transformación y/o recuperación de células cuya estandarización de estos métodos va más allá de los objetivos de esta tesis y no fueron explorados. Sin embargo, aún cuando este análisis no arrojó resultados experimentales concluyentes, se proponen otros métodos potencialmente útiles para estudios posteriores.

Bibliografía

- Acencio, M. y Lemke, N., (2009). Towards the prediction of essential genes by integration of network topology, cellular localization and biological process information. *BMC Bioinformatics*, Volumen 10, p. 290.
- Aittokallio, T. y Schwikowski, B., (2006). Graph-based methods for analysing networks in cell biology. *Briefings in Bioinformatics*, 7(3), pp. 243-255.
- Alberts, B., (1998). The cell as a collection of protein machines: preparing the next generation of molecular biologists. *Cell*, 92(3), p. 291-4.
- Al-Shahrour, F., Díaz-Uriarte, R. y Dopazo, J., (2004). FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes. *BMC Bioinformatics*, 20(4), p. 578-80.
- Altschul, S., et al., (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, 25(17), pp. 3389-402.
- Altschul, S., et al., (2005). Protein database searches using compositionally adjusted substitution matrices. *FEBS Journal*, 272(20), pp. 5101-9.
- Ashburner, M., et al., (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.. *Nature Genetics*, 1(25), pp. 25-9.
- Baker, M., (2013). Big biology: The 'omes puzzle. *Nature*, 494(7438), pp. 416-9.
- Ball, P., (2013). DNA: Celebrate the unknowns. *Nature*, 496(7446), pp. 419-20.
- Becker, R., Chambers, J. y Wilks, A., (1988). *The New S Language*. New York: Chapman & Hall.
- Benkovic, S., Theriot, J. y Ringe, D., (2013). Open questions - in brief: Beyond -omics, missing motor proteins, and getting from molecules to organisms. *BMC Biology*, 11(8).
- Blank, L., Kuepfer, L. y Sauer, U., (2005). Large-scale 13C-flux analysis reveals mechanistic principles of metabolic network robustness to null mutations in yeast. *Genome Biology*, 6(6), p. R49.
- Bouteldja, N. y Timson, D., (2010). The biochemical basis of hereditary fructose intolerance. *Journal of Inherited Metabolic Diseases*, 33(2), pp. 105-12.
- Burke, Dawson y Stearns, (2005). "Quick and Dirty" plasmid transformation of yeast colonies. En: *Methods in Yeast Genetics*. Plainview, NY: CSHL Press.
- Burke, Dawson y Stearns, (2005). Yeast Colony PCR. En: *Methods in Yeast Genetics*. Plainview, NY: CSHL Press.
- Burke, Dawson y Stearns, (2005). Yeast Sporulation. En: *Methods in Yeast Genetics*. Plainview, NY: CSHL Press.
- Camon, E., et al., (2004). The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology. *Nucleic Acids Research*, 32(Database issue), pp. D262-6.
- Castrillo, J. y Oliver, S., (2011). Yeast Systems Biology: The Challenge of Eukaryotic Complexity. *Methods in Molecular Biology*, Volumen 759, pp. 3-28.
- Chalker, A. y Lunsford, R., (2002). Rational identification of new antibacterial drug targets that are essential for viability using a genomics-based approach. *Pharmacology & Therapeutics*, 95(1), pp. 1-20.
- Chen, D., Yang, B. y Kuo, T., (1992). One-step transformation of yeast in stationary phase. *Current Genetics*, 21(1), pp. 83-4.
- Chen, X., Wang, X., Kaufman, B. y Butow, R., (2005). Aconitase couples metabolic regulation to mitochondrial DNA maintenance. *Science*, 307(5710), pp. 714-7.

- Cherry, J., et al., (2012). Saccharomyces Genome Database: the genomics resource of budding yeast. *Nucleic Acids Research*, 40(Database issue), pp. D700-5.
- Chuang, H., Hofree, M. y Ideker, T., (2010). A Decade of Systems Biology. *Annual Review of Cell and Developmental Biology*, Volumen 26, p. 721–44.
- Compagno, C., Boschi, F. y Ranzi, B., (1996). Glycerol production in a triose phosphate isomerase deficient mutant of *Saccharomyces cerevisiae*. *Biotechnology Progress*, 12(5), pp. 591-5.
- Compagno, C., et al., (2001). Alterations of the glucose metabolism in a triose phosphate isomerase-negative *Saccharomyces cerevisiae* mutant. *Yeast*, 18(7), pp. 663-70.
- Cordier, H., Mendes, F., Vasconcelos, I. y François, J., (2007). A metabolic and genomic study of engineered *Saccharomyces cerevisiae* strains for high glycerol production. *Metabolic engineering*, 9(4), pp. 364-78.
- Costanzo, M., (2010). The Genetic Landscape of a Cell. *Science*, 327(5964), pp. 425-31.
- Cusack, M., Thibert, B., Bredesen, D. y Del Rio, G., (2007). Efficient identification of critical residues based only on protein structure by network analysis. *PLoS One*, 2(5), p. e421.
- Cusick, M., Klitgord, N., Vidal, M. y Hill, D., (2005). Interactome: gateway into systems biology. *Human Molecular Genetics*, 14(2), pp. R171-81.
- Dada, J. y Mendes, P., (2011). Multi-scale modelling and simulation in systems biology. *Integrative Biology : quantitative biosciences from nano to macro*, 3(2), p. 86–96.
- de Matos Simoes, R., Dehmer, M. y Emmert-Streib, F., (2013). Interfacing cellular networks of *S. cerevisiae* and *E. coli*: Connecting dynamic and genetic information. *BMC Genomics*, Issue 14, p. 324.
- del Rio, G., Koschützki, D. y Coello, G., (2009). How to identify essential genes from molecular networks?. *BMC Systems Biology*, Issue 3, p. 102.
- Delaunay, A., et al., (2002). A thiol peroxidase is an H₂O₂ receptor and redox-transducer in gene activation. *Cell*, 111(4), pp. 471-81.
- Dezso, Z., Oltvai, Z. y Barabási, A., (2003). Bioinformatics analysis of experimentally determined protein complexes in the yeast *Saccharomyces cerevisiae*. *Genome Research*, 13(11), pp. 2450-4.
- Dixit, P. y Maslov, S., (2013). Evolutionary Capacitance and Control of Protein Stability in Protein-Protein Interaction Networks. *PLoS Comput Biol*, 9(4), p. e1003023.
- Dolinski, K., Chatr-Aryamontri, A. y Tyers, M., (2013). Systematic curation of protein and genetic interaction data for computable biology. *BMC Biology*, Issue 11, p. 43.
- Duarte, N., Herrgård, M. y Palsson, B., (2004). Reconstruction and validation of *Saccharomyces cerevisiae* metabolic network iND750, a fully compartmentalized genome-scale metabolic model. *Genome Research*, 14(7), pp. 1298-309.
- Dunn, B., et al., (2012). Analysis of the *Saccharomyces cerevisiae* pan-genome reveals a pool of copy number variants distributed in diverse yeast strains from differing industrial environments. *Genome Research*, 22(5), pp. 908-24.
- Ejiri, S., (2002). Moonlighting functions of polypeptide elongation factor 1: from actin bundling to zinc finger protein R1-associated nuclear localization. *Bioscience, biotechnology, and biochemistry*, 66(1), pp. 1-21.
- Estrada, E., (2006). Virtual identification of essential proteins within the protein interaction network of yeast. *Proteomics*, 6(1), p. 35–40.
- Fernández, A. y Lynch, M., (2011). Non-adaptive origins of interactome complexity. *Nature*, 474(7352), pp. 502-5.
- Förster, J., et al., (2003). Genome-scale reconstruction of the *Saccharomyces cerevisiae* metabolic network. *Genome Research*, 13(2), pp. 244-253.

- Förster, J., Famili, I., Palsson, B. y Nielsen, J., (2003). Large-scale evaluation of in silico gene deletions in *Saccharomyces cerevisiae*. *OMICS*, 7(2), pp. 193-202.
- Friedel, C. y Zimmer, R., (2006). Inferring topology from clustering coefficients in protein-protein interaction networks. *BMC Bioinformatics*, Issue 7, p. 519.
- Gancedo, C. y Flores, C., (2008). Moonlighting Proteins in Yeasts. *Microbiology and Molecular Biology Reviews*, 72(1), pp. 197-210.
- Garí, E., Piedrafita, L., Aldea, M. y Herrero, E., (1997). A Set of Vectors with a Tetracycline-Regulatable Promoter System for Modulated Gene Expression in *Saccharomyces cerevisiae*. *YEAST*, 13(9), p. 837-48.
- Gavin, A., et al., (2006). Proteome survey reveals modularity of the yeast cell machinery. *Nature*, 440(7084), p. 631-6.
- Gebert, N., et al., (2011). Dual function of Sdh3 in the respiratory chain and TIM22 protein translocase of the mitochondrial inner membrane. *Molecular cell*, 44(5), p. 811-8.
- Georgi, B., Voight, B. y Bucán, M., (2013). From Mouse to Human: Evolutionary Genomics Analysis of Human Orthologs of Essential Genes. *PLoS Genet*, 9(5), p. e1003484.
- Giaever, G., et al., (2002). Functional profiling of the *Saccharomyces cerevisiae* genome. *Nature*, 418(6896), p. 387-91.
- Gietz, R. y Woods, R., (2002). Transformation of Yeast by the Liac/SS Carrier DNA/PEG Method. *Methods in Enzymology*, Volumen 350, pp. 87-96.
- Gillespie, D., (1976). A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *Journal of Computational Physics*, 22(4), p. 403-34.
- Gómez, A., et al., (2003). Do current sequence analysis algorithms disclose multifunctional (moonlighting) proteins?. *Bioinformatics*, 19(7), pp. 895-6.
- Go, M., Koudelka, A., Amyes, T. y Richard, J., (2010). Role of Lys-12 in Catalysis by Triosephosphate Isomerase: A Two-Part Substrate Approach. *Biochemistry*, 49(25), p. 5377-89.
- Gonzalez, F., Delahodde, A., Kodadek, T. y Johnston, S., (2002). Recruitment of a 19S proteasome subcomplex to an activated promoter. *Science*, 296(5567), pp. 548-50.
- Gueldener, U., et al., (2002). A second set of loxP marker cassettes for Cre-mediated multiple gene knockouts in budding yeast. *Nucleic Acids Research*, 30(6), p. e23.
- Hahn, M. y Kern, A., (2005). Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks. *Molecular Biology and Evolution*, 22(4), pp. 803-6.
- Hall, D., et al., (2004). Regulation of gene expression by a metabolic enzyme. *Science*, 306(5695), pp. 482-4.
- Han, J., et al., (2004). Evidence for dynamically organized modularity in the yeast protein-protein interaction network. *Nature*, 430(6995), pp. 88-93.
- Hanley, J. y McNeil, B., (1982). The meaning and use of the area under the receiver characteristic (ROC) curve. *Radiology*, 143(1), pp. 29-36.
- Hart, G., Lee, I. y Marcotte, E., (2007). A high-accuracy consensus map of yeast protein complexes reveals modular nature of gene essentiality. *BMC Bioinformatics*, Issue 8, p. 236.
- Hartwell, L., Hopfield, J., Leibler, S. y Murray, A., (1999). From molecular to modular cell biology. *Nature*, 402(6761 Suppl), p. C47-52.
- Hasegawa, H. y Holm, L., (2009). Advances and pitfalls of protein structural alignment. *Current Opinion in Structural Biology*, 19(3), pp. 341-8.
- Hegemann, J. y Heick, S., (2011). Delete and repeat: a comprehensive toolkit for sequential gene knockout in the budding yeast *Saccharomyces cerevisiae*. *Methods in Molecular Biology*, Issue 765, pp. 189-206.

- Heinemann, M. y Sauer, U., (2010). Systems biology of microbial metabolism. *Current Opinion in Microbiology*, 13(3), p. 337–43.
- He, X. y Zhang, J., (2006). Why Do Hubs Tend to Be Essential in Protein Networks?. *PLoS Genetics*, 2(6), p. e88.
- Ho, C., et al., (2009). A molecular barcoded yeast ORF library enables mode-of-action analysis of bioactive compounds. *Nature Biotechnology*, 27(4), pp. 369-77.
- Hormozdiari, F., Berenbrink, P., Pržulj, N. y Sahinalp, S. C., (2007). Not all scale-free networks are born equal: the role of the seed graph in PPI network evolution. *PLoS Computational Biology*, 3(7), p. e118.
- Hübner, K., Sahle, S. y Kummer, U., (2011). Applications and trends in systems biology in biochemistry. *FEBS Journal*, 278(16), p. 2767–2857.
- Ito, H., Fukuda, K., Murata, K. y Kimura, A., (1983). Transformation of intact cells treated with alkali cations. *Journal of Bacteriology*, 153(1), pp. 163-8.
- Ito, T., et al., (2001). A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proceedings of the National Academy of Science*, 98(8), p. 4569–74.
- Janga, S., Díaz-Mejía, J. y Moreno-Hagelsieb, G., (2010). Network-based function prediction and interactomics: The case for metabolic enzymes. *Metabolic engineering*, 13(1), pp. 1-10.
- Jang, H., et al., (2004). Two enzymes in one; two yeast peroxiredoxins display oxidative stress-dependent switching from a peroxidase to a molecular chaperone function. *Cell*, 117(5), pp. 625-35.
- Jeffery, C., (1999). Moonlighting proteins. *Trends in Biochemical Sciences*, 24(1), pp. 8-11.
- Jeffery, C., (2003). Moonlighting proteins: old proteins learning new tricks. *Trends in Genetics*, 18(8), p. 415–7.
- Jeffery, C., (2004). Molecular mechanisms for multitasking: recent crystal structures of moonlighting proteins. *Current Opinion in Structural Biology*, 14(6), pp. 663-8.
- Jeong, H., Mason, S., Barabási, A. y Oltvai, Z., (2001). Lethality and centrality in protein networks. *Nature*, 411(6833), pp. 41-2.
- Johnson, M. y Hummer, G., (2013). Interface-Resolved Network of Protein-Protein Interactions. *PLoS Computational Biology*, 9(5), p. e1003065.
- Kaltenbach, H. y Stelling, J., (2012). Modular analysis of biological networks. *Advances in Experimental Medicine and Biology*, Issue 736, pp. 3-17.
- Kaspar, B., Bifano, A. y Caprara, M., (2008). A shared RNA-binding site in the Pet54 protein is required for translational activation and group I intron splicing in yeast mitochondria. *Nucleic acids research*, 36(9), pp. 2958-68.
- Kent, E., Hoops, S. y Mendes, P., (2012). Condor-COPASI: high-throughput computing for biochemical networks. *BMC Systems Biology*, Issue 6, p. 91.
- Khurana, E., Fu, Y., Chen, J. y Gerstein, M., (2013). Interpretation of Genomic Variants Using a Unified Biological Network Approach. *PLoS Comput Biol*, 9(3), p. e1002886.
- Kim, D., et al., (2010). Analysis of a genome-wide set of gene deletions in the fission yeast *Schizosaccharomyces pombe*. *Nature Biotechnology*, 28(6), p. 617.
- Kim, P., Lu, L., Xia, Y. y Gerstein, M., (2006). Relating three-dimensional structures to protein networks provides evolutionary insights. *Science*, 314(5807), pp. 1938-41.
- Klebe, R., Harris, J., Sharp, D. y Douglas, M., (1983). A general method for polyethylene-glycol-induced genetic transformation of bacteria and yeast. *Gene*, 25(2-3), p. 333–41.
- Klipp, E., (2007). Modelling dynamic processes in yeast. *Yeast*, 24(11), pp. 943-59.
- Knüpfer, C., Beckstein, C., Dittrich, P. y Le Novère, N., (2013). Structure, function, and behaviour of computational models in systems biology. *BMC Systems Biology*, Issue 7, p. 43.

- Kohlstedt, M., Becker, J. y Wittmann, C., (2010). Metabolic fluxes and beyond—systems biology understanding and engineering of microbial metabolism. *Applied Microbiology and Biotechnology*, 88(5), p. 1065–75.
- Komives, E., et al., (1991). Electrophilic catalysis in triosephosphate isomerase: the role of histidine-95. *Biochemistry*, 30(12), p. 3011–9.
- Krogan, N., et al., (2006). Global landscape of protein complexes in the yeast *Saccharomyces cerevisiae*. *Nature*, 440(7084), p. 637–43.
- Lander, A., (2010). The edges of understanding. *BMC Biology*, Issue 8, p. 40.
- Leek, J., Taub, M. y Rasgon, J., (2012). A statistical approach to selecting and confirming validation targets in -omics experiments. *BMC Bioinformatics*, Issue 13, p. 150.
- Lewis, A., Jones, N., Porter, M. y Deane, C., (2012). What Evidence Is There for the Homology of Protein-Protein Interactions?. *PLoS Comput Biol*, 8(9), p. e1002645.
- Li, M. y Elledge, S., (2005). MAGIC, an in vivo genetic method for the rapid construction of recombinant DNA molecules. *Nature Genetics*, 37(3), pp. 311-9.
- Lim, M., et al., (2002). On the release of cytochrome c from mitochondria during cell death signaling. *Journal of biomedical science*, 9(6 Pt 1), pp. 488-506.
- Lin, A., McCammon, M. y McAlister-Henn, L., (2001). Kinetic and physiological effects of alterations in homologous isocitrate-binding sites of yeast NAD(+)-specific isocitrate dehydrogenase. *Biochemistry*, 40(47), pp. 14291-301.
- Lindgren, G., Hwang, Y., Oshima, Y. y Lindgren, C., (1965). Genetical mutants induced by ethyl methanesulfonate in *Saccharomyces*. *Canadian Journal of Genetics and Cytology*, 7(3), p. 491–9.
- Liu, Y., Devescovi, V., Chen, S. y Nardini, C., (2013). Multilevel omic data integration in cancer cell lines: advanced annotation and emergent properties. *BMC Systems Biology*, Issue 7, p. 14.
- Lobo, Z., (1984). *Saccharomyces cerevisiae* aldolase mutants. *Journal of Bacteriology*, 1(160), pp. 222-6.
- Lu, M., et al., (2001). Interaction between aldolase and vacuolar H⁺-ATPase: evidence for direct coupling of glycolysis to the ATP-hydrolyzing proton pump. *The Journal of biological chemistry*, 276(32), pp. 30407-13.
- Lynch, M., (2011). The evolution of multimeric protein assemblages. *Molecular Biology and Evolution*, 29(5), p. 1353–66.
- Ma, B., Tsai, C., Pan, Y. y Nussinov, R., (2010). Why does binding of proteins to DNA or proteins to proteins not necessarily spell function?. *ACS Chemical Biology*, 5(3), pp. 265-72.
- Ma, H. y Zeng, A., (2003). The connectivity structure, giant strong component and centrality of metabolic networks. *Bioinformatics*, 19(11), pp. 1423-30.
- Masino, L., et al., (2011). Functional interactions as a survival strategy against abnormal aggregation. *The FASEB journal*, 25(1), p. 45–54.
- Medina, I., et al., (2010). Babelomics: an integrative platform for the analysis of transcriptomics, proteomics and genomic data with advanced functional profiling. *Nucleic Acids Research*, 38(Web Server issue), pp. W210-3.
- Mesarović, M. y Takahara, Y., (1972). Multilevel, hierarchical, systems theory. En: s.l.:Academic press.
- Metcalf, W., Jiang, W. y BL, W., (1994). Use of the rep technique for allele replacement to construct new *Escherichia coli* hosts for maintenance of R6Kλ origin plasmids at different copy numbers. *Gene*, 138(1-2), pp. 1-7.
- Mnaimneh, S., et al., (2004). Exploration of essential gene functions via titratable promoter alleles. *Cell*, 118(1), pp. 31-44.
- Newman, M., (2006). Modularity and community structure in networks. *Proceedings of the National Academy of Sciences of the United States of America*, 103(23), p. 8577–82.

- Nikolaev, Y., et al., (2010). The Leucine Zipper Domains of the Transcription Factors GCN4 and c-Jun Have Ribonuclease Activity. *PLoS ONE*, 5(5), p. e10765.
- Ogata, H., et al., (1999). KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research*, 27(1), pp. 29-34.
- Ogmen, U., et al., (2005). PRISM: protein interactions by structural matching. *Nucleic Acids Research*, 33(Web Server issue), pp. W331-6.
- Ooi, S., Shoemaker, D. y Boeke, J., (2003). DNA helicase gene interaction network defined using synthetic lethality analyzed by microarray. *Nature Genetics*, 35(3), pp. 277-86.
- Orosz, F., Olah, J. y Ovadi, J., (2009). Triosephosphate isomerase deficiency: new insights into an enigmatic disease. *Biochimica et Biophysica Acta*, 1792(12), pp. 1168-74.
- Österlund, T., Nookaew, I., Bordel, S. y Nielsen, J., (2013). Mapping condition-dependent regulation of metabolism in yeast through genome-scale modeling. *BMC Systems Biology*, Issue 7, p. 36.
- Otto, A., et al., (1986). Kinetic effects of fructose 1,6-bisphosphate on yeast phosphofructokinase. *Biomedica biochimica acta*, 45(7), pp. 865-75.
- Overkamp, K., et al., (2002). Metabolic Engineering of Glycerol Production in *Sacharomyces cerevisiae*. *Applied and Environmental Microbiology*, 68(6), pp. 2814-21.
- Oyedotun, K. y Lemire, B., (2004). The quaternary structure of the *Sacharomyces cerevisiae* succinate dehydrogenase. Homology modeling, cofactor docking, and molecular dynamics simulation studies. *The Journal of Biological Chemistry*, 279(10), pp. 9424-31.
- Pagel, P., et al., (2005). The MIPS mammalian protein-protein interaction database. *Bioinformatics*, 21(6), pp. 832-4.
- Palsson, B., (2009). Metabolic systems biology. *FEBS Letters*, 583(24), p. 3900-4.
- Papp, B., Pál, C. y Hurst, L., (2004). Metabolic network analysis of the causes and evolution of enzyme dispensability in yeast. *Nature*, 429(6992), p. 661-4.
- Pawlowski, P., Kaczanowski, S. y Zielenkiewicz, P., (2013). A kinetic model of the evolution of a protein interaction network. *BMC Genomics*, Volumen 14, p. 172.
- Payne, J. y Wagner, A., (2013). Constraint and Contingency in Multifunctional Gene Regulatory Circuits. *PLoS Computational Biology*, 9(6), p. e1003071.
- Pechmann, S., Levy, E., Tartaglia, G. y Vendruscolo, M., (2009). Physicochemical principles that regulate the competition between functional and dysfunctional association of proteins. *Proceedings of the National Academy of Science*, 106(25), p. 10159-64.
- Pereira-Leal, J., Levy, E. y Teichmann, S., (2006). The origins and evolution of functional modules: lessons from protein complexes. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 361(1467), p. 507-17.
- Petersen, J. y Holmberg, S., (1986). The ILV5 gene of *Saccharomyces cerevisiae* is highly expressed. *Nucleic Acids Research*, 14(24), pp. 9631-51.
- Petranovic, D., Tyo, K., Vemuri, G. y Nielsen, J., (2010). Prospects of yeast systems biology for human health: integrating lipid, protein and energy metabolism. *FEMS Yeast Research*, 10(8), p. 1046-59.
- Piatigorsky, J., et al., (1988). Gene sharing by delta-crystallin and argininosuccinate lyase. *Proceedings of the National Academy of Science U S A*, 85(10), pp. 3479-83.
- Pržulj, N., Corneil, D. G. y Jurisica, I., (2004). Modeling interactome: scale-free or geometric?. *Bioinformatics*, 20(18), p. 3508-15.
- Punta, M., et al., (2012). The Pfam protein families database. *Nucleic Acids Research*, 40(Database Issue), pp. D290-301.
- Ralser, M., et al., (2006). Triose phosphate isomerase deficiency is caused by altered dimerization--not catalytic inactivity--of the mutant enzymes. *PLoS One*, Issue 1, p. e30.

- Reguly, T., et al., (2006). Comprehensive curation and analysis of global interaction networks in *Saccharomyces cerevisiae*. *Journal of Biology*, 5(4), p. 11.
- Reznik, E., Yohe, S. y Segrè, D., (2013). Invariance and optimality in the regulation of an enzyme. *Biology Direct*, Issue 8, p. 7.
- Rieder, S. y Rose, I., (1959). The Mechanism of the Triosephosphate Isomerase Reaction. *The Journal of Biological Chemistry*, 234(5), pp. 1007-10.
- Rohwer, J., Schuster, S. y Westerhoff, H., (1996). How to recognize monofunctional units in a metabolic system. *Journal of Theoretical Biology*, 179(3), pp. 213-28.
- Roland, B., et al., (2013). Evidence of a triosephosphate isomerase non-catalytic function critical to behavior and longevity. *Journal of Cell Science*, 126(Pt 14), pp. 3151-8.
- Rosslenbroich, B., (2011). Outline of a concept for organismic systems biology. *Seminars in Cancer Biology*, 21(3), p. 156– 64.
- Rothstein, R., (1991). Targeting, disruption, replacement, and allele rescue: integrative DNA transformation in yeast. *Methods in Enzymology*, Volumen 194, pp. 281-301.
- Ruiz, A., et al., (2009). Moonlighting proteins Hal3 and Vhs3 form a heteromeric PPCDC with Ykl088w in yeast CoA biosynthesis. *Nature chemical biology*, 5(12), pp. 920-8.
- Rutherford, S., Hirate, Y. y Swalla, B., (2007). The Hsp90 Capacitor, Developmental Remodeling, and Evolution: The Robustness of Gene Networks and the Curious Evolvability of Metamorphosis. *Critical Reviews in Biochemistry and Molecular Bioogy*, 42(5), p. 355–72.
- Rutherford, S. y Lindquist, S., (1998). Hsp90 as a capacitor for morphological evolution. *Nature*, 396(6709), p. 336–42.
- Ryan, C., Krogan, N., Cunningham, P. y Cagney, G., (2013). All or nothing: protein complexes flip essentiality between distantly related eukaryotes. *Genome Biology and Evolution*, 5(6), pp. 1049-59.
- Saha, S. y Heber, S., (2006). In silico prediction of yeast deletion phenotypes. *Genetics and Molecular Research*, 5(1), pp. 224-32.
- Samanta, M., Murthy, M., Balaram, H. y Balaram, P., (2011). Revisiting the Mechanism of the Triosephosphate Isomerase Reaction: The Role of the Fully Conserved Glutamic Acid 97 Residue. *ChemBioChem*, 12(12), pp. 1886-96.
- Schnoes, A., Brown, S., I, D. y Babbitt, P., (2009). Annotation error in public databases: Misannotation of molecular function in enzyme superfamilies. *PLoS Computational Biology*, 5(12), p. e1000605.
- Schnoes, A., et al., (2013). Biases in the Experimental Annotations of Protein Function and Their Effect on Our Understanding of Protein Function Space. *PLoS Computational Biology*, 9(5), p. e1003063.
- Schuldiner, M., et al., (2005). Exploration of the function and organization of the yeast early secretory pathway through an epistatic miniarray profile.. *Cell*, 123(3), pp. 507-19.
- Segel, I., (1975). *Enzyme kinetics: behavior and analysis of rapid equilibrium and steady-state enzyme systems*. New York: Wiley.
- Semple, J., Vavouri, T. y Lehner, B., (2008). A simple principle concerning the robustness of protein complex activity to changes in gene expression. *BMC Systems Biology*, Issue 2, p. 1.
- Sharan, R. y Ideker, T., (2006). Modeling cellular machinery through biological network comparison. *Nature Biotechnology*, 24(4), pp. 427-33.
- Shi, Y., et al., (2005). Genetic perturbation of glycolysis results in inhibition of de novo inositol biosynthesis. *The Journal of Biological Chemistry*, 280(51), pp. 41805-10.
- Skarnes, W., et al., (2011). A conditional knockout resource for the genome-wide study of mouse gene function. *Nature*, 474(7351), p. 337–42.
- Song, J. y Singh, M., (2013). From Hub Proteins to Hub Modules: The Relationship Between Essentiality and Centrality in the Yeast Interactome at Different Scales of Organization. *PLoS Comput Biol*, 9(2), p. e1002910.

- Southern, J., et al., (2008). Multi-scale computational modelling in biology and physiology. *Progress in Biophysics and Molecular Biology*, 96(1-3), p. 60–89.
- Sriram, G., et al., (2005). Single-Gene Disorders: What Role Could Moonlighting Enzymes Play?. *American Journal of Human Genetics*, 76(6), p. 911–24.
- St Onge, R., et al., (2007). Systematic pathway analysis using high-resolution fitness profiling of combinatorial gene deletions. *Nature Genetics*, 39(2), pp. 199-206.
- Strohmann, R., (1993). Ancient genomes, wise bodies, unhealthy people: limits of a genetic paradigm in biology and medicine. *Perspectives in Biology and Medicine*, 37(1), pp. 112-45.
- Suzuki, C., et al., (1997). ATP-dependent proteases that also chaperone protein biogenesis. *Trends in biochemical sciences*, 22(4), pp. 118-23.
- Tchórzewski, M., Boldyreff, B. y Grankowski, N., (1999). Extraribosomal function of the acidic ribosomal P1-protein YP1alpha from *Saccharomyces cerevisiae*. *Acta biochimica Polonica*, 46(4), pp. 901-10.
- Teusink, B. y Westerhoff, H., (2000). 'Slave' metabolites and enzymes. A rapid way of delineating metabolic control. *European Journal of Biochemistry*, 267(7), pp. 1889-93.
- The Gene Ontology Consortium, (2004). The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Research*, 32(Database issue), p. D258–D261.
- Thellin, O., et al., (1999). Housekeeping genes as internal standards: use and limits. *Journal of Biotechnology*, 75(2-3), pp. 291-5.
- Thibert, B., Bredesen, D. y del Rio, G., (2005). Improved prediction of critical residues for protein function based on network and phylogenetic analyses. *BMC Bioinformatics*, Issue 6, p. 213.
- Tompa, P., Szász, C. y Buday, L., (2005). Structural disorder throws new light on moonlighting. *Trends in biochemical sciences*, 30(9), pp. 484-9.
- Tong, A., et al., (2001). Systematic genetic analysis with ordered arrays of yeast deletion mutants. *Science*, 294(5550), pp. 2364-8.
- Tsai, C., Ma, B. y Nussinov, R., (2009). Protein-protein interaction networks: how can a hub protein bind so many different partners?. *Trends in Biochemical Sciences*, 34(12), pp. 594-600.
- Tung, H., Wang, W. y Chan, C., (1995). Regulation of chromosome segregation by Glc8p, a structural homolog of mammalian inhibitor 2 that functions as both an activator and an inhibitor of yeast protein phosphatase 1. *Molecular and cellular biology*, 15(11), pp. 6064-74.
- Twycross, J., et al., (2010). Stochastic and deterministic multiscale models for systems biology: an auxin-transport case study. *BMC Systems Biology*, Issue 4, p. 34.
- Uetz, P., et al., (2000). A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. *Nature*, 403(6770), pp. 623-631.
- UniProt Consortium, (2012). Reorganizing the protein space at the Universal Protein Resource (UniProt). *Nucleic Acids Research*, 40(Database issue), pp. D71-75.
- Varki, A., (2013). Omics: Account for the 'dark matter' of biology. *Nature*, 497(7451), p. 565.
- Vidal, M., (2005). Interactome modeling. *FEBS Letters*, 579(8), p. 1834–8.
- Vitkup, D., Kharchenko, P. y Wagner, A., (2006). Influence of metabolic network structure and function on enzyme evolution. *Genome Biology*, 7(5), p. R39.
- von Bertalanffy, L., (1976). *General System Theory: Foundations, Development, Applications*. Revised edition ed. New York: Geroge Braziller, Inc..
- Wagner, A., (2000). Robustness against mutations in genetic networks of yeast. *Nature Genetics*, 24(4), p. 355–61.

- Walker, D. y Southgate, J., (2009). The virtual cell—a candidate co-ordinator for 'middle-out' modelling of biological systems. *Briefings in Bioinformatics*, 10(4), p. 450–61.
- Wang, H., et al., (2009). A complex-based reconstruction of the *Saccharomyces cerevisiae* interactome. *Molecular & cellular proteomics: MCP*, 8(6), p. 1361–81.
- Weiss, P., (1969). The living system: determinism stratified. En: A. Koestler y J. Smythies, edits. *Beyond reductionism. New perspectives in the life sciences*. s.l.:s.n.
- Wellstead, P., et al., (2008). The rôle of control and system theory in systems biology. *Annual Reviews in Control*, 32(1), p. 33–47.
- West, G., (2013). Wisdom in Numbers. *Scientific American*, 308(14).
- Winzeler, E., et al., (1999). Funtional characterization of the *S. cerevisiae* genome by gene deletion and parallel analysis. *Science*, 285(5429), pp. 901-6.
- Woese, C., (2004). A new biology for a new century. *Microbiology and Molecular Biology Reviews: MMBR*, 68(2), p. 137–86.
- Wunderlich, Z. y Mirny, L., (2006). Using the topology of metabolic networks to predict viability of mutant strains. *Biophysics Journal*, 91(6), pp. 2304-11.
- Yan Tong, A. H. y Boone, C., (2005). *Yeast Protocols. Running head: Synthetic Genetic Array (SGA) Analysis*. Second Edition ed. Totowa, NJ, U. S. A.: Methods in Molecular Biology, The Humana Press Inc..
- Yu, H., et al., (2007). The importance of bottlenecks in protein networks: correlation with gene essentiality and expression dynamics. *PLoS Computational Biology*, 3(4), p. e59.
- Zelenaya-Troitskaya, O., Perlman, P. y Butow, R., (1995). An enzyme in yeast mitochondria that catalyzes a step in branched-chain amino acid biosynthesis also functions in mitochondrial DNA stability. *EMBO Journal*, 14(13), p. 3268–76.
- Zhang, Q., et al., (2012). Structure-based prediction of protein-protein interactions on a genome-wide scale. *Nature*, 490(7421), pp. 556-60.
- Zhu, W., et al., (2002). Evidence that the pre-mRNA splicing factor Clf1p plays a role in DNA replication in *Saccharomyces cerevisiae*. *Genetics*, 160(4), pp. 1319-33.
- Zotenko, E., Mestre, J., O'Leary, D. y Przytycka, T., (2008). Why Do Hubs in the Yeast Protein Interaction Network Tend To Be Essential?: Reexamining the Connection between the Network Topology and Essentiality. *PLoS Computational Biology*, 4(8), p. e1000140.