



**CENTRO DE INVESTIGACIÓN Y ESTUDIOS
AVANZADOS DEL INSTITUTO POLITÉCNICO
NACIONAL**

UNIDAD IRAPUATO

Estudio sistemático de la divergencia
evolutiva y funcional de la duplicación
completa del genoma en *Saccharomyces
cerevisiae*

Tesis que presenta

M. en C. Diana Itzel Ascencio Sánchez

Para obtener el grado de

Doctor en Ciencias

En la especialidad de

Biotecnología de Plantas

Director de Tesis: Dr. Alexander de Luna Fors

Irapuato, Guanajuato

Enero del 2016

Este trabajo fue realizado en el *Laboratorio de Sistemas Genéticos* en la *Unidad de Genómica Avanzada del Cinvestav Irapuato* bajo la dirección del Dr. Alexander de Luna Fors. Dicho trabajo fue financiado con una beca de doctorado (No. Becario: 267797) y por el proyecto CB2011/164889 del Consejo Nacional de Ciencia y Tecnología (Conacyt)

AGRADECIMIENTOS

Al Dr. Alexander de Luna Fors por ser un gran tutor y mentor – creativo, crítico, perfeccionista pero paciente. Muchas gracias por guiarme y apoyarme en este largo viaje.

A mi comité tutorial: Dra. Riego, Dra. Simpson, Dr. Gillmor y el Dr. Barona por todo el tiempo dedicado a ayudarme y por exigirme siempre lo mejor.

A todos los que son y fueron miembros del laboratorio de Sistemas Genéticos: Citlali, Ana Lilia, Emiliano, Paulina, Luis, Adrián, Erika, Jorge, Jonathan, Adriana, Abraham, Sergio, Alex J. y otros integrantes que pasaron por nuestro laboratorio. He aprendido muchas cosas platicando y discutiendo con cada uno de ustedes tanto de la ciencia como de la vida. Les agradezco siempre haber creado un ambiente agradable de trabajo.

A nuestros colaboradores: Dr. Luis Delaye, Dr. Christian C. R. Landry, Soledad Ochoa y Guillaume Diss por su invaluable contribución a este proyecto.

Al CONACYT por proporcionarme becas (No. de registro: **267797**) de doctorado y beca mixta, este financiamiento me permitió dedicarme completamente a este trabajo. Las becas de posgrado realmente ayudan al desarrollo de la ciencia que es indispensable en un país como el nuestro.

A Rosy por todo el cariño, soporte y felicidad que me ha brindado durante todos estos años.

Finalmente, me gustaría dedicar este trabajo a mis padres que son una inspiración y un ejemplo de vida. Siempre me han apoyado incondicionalmente en todo lo que he querido hacer. Muchas gracias.

ÍNDICE GENERAL

Resumen	1
Abstract	3
Introducción	4
Objetivos	13
Capítulo 1: Estudio de la dinámica evolutiva de genes parálogos después de una Duplicación de Genoma Completo	14
1.1 Antecedentes.....	14
1.2 Hipótesis.....	16
1.3 Métodos.....	16
1.3.1 Selección de ortogrupos, secuencias, árboles filogenéticos y alineamientos usados.....	16
1.3.2 Pruebas filogenéticas con máxima verosimilitud con CodeML.....	18
1.3.3 Grupos funcionales, abundancia, adecuación y desorden intrínseco	20
1.4 Resultados	22
1.4.1 La mayor parte de los ortogrupos experimentan un cambio en la tasa de evolución después de la WGDs.....	22
1.4.2 Las proteínas con funciones diferentes muestran tendencias evolutivas diferentes	24
1.4.3 La abundancia de proteína, el desorden intrínseco y la contribución a la adecuación son factores determinantes en el cambio de las tasas de evolución después de la WGD.....	27
1.4.4 Una buena parte de los ortogrupos experimenta una aceleración fuerte inmediatamente después de la duplicación seguida por un periodo de estabilización.....	30
1.4.5 Algunos ortogrupos muestran cambios asimétricos en sus tasas de evolución después de la duplicación	31
1.5 Discusión.....	35
Capítulo 2: Divergencia funcional de las regiones <i>cis</i>-regulatoria y codificante de genes duplicados en levadura	38
2.1 Antecedentes.....	39
2.2 Hipótesis.....	40
2.3 Métodos.....	41
2.3.1 Cepas y plásmidos usados.....	41

2.3.3	Construcción de cepas quiméricas y controles.....	43
2.3.4	Calibración de condiciones usadas en la competencia.....	45
2.3.5	Construcción de cepas de promotor regulable <i>tetO₇</i>	47
2.3.6	Método de competencia automatizado.....	47
2.4	Resultados	48
2.4.1	Un 10% de las parejas de duplicados WGD presentan una divergencia fenotípica asimétrica.....	48
2.4.2	Comparación directa de la divergencia fenotípica de la región <i>cis</i> -regulatoria y codificante de genes parálogos	51
2.4.3	Las parejas de duplicados con divergencia funcional extrema exhiben divergencia total en la región codificante y <i>cis</i> -regulatoria.....	58
2.4.4	Nuestro ensayo con promotores regulables indica que la dosis génica de los parálogos estudiados afecta fuertemente la adecuación	61
2.5	Discusión.....	64
	Discusión general.....	67
	Conclusiones	70
	Preguntas abiertas.....	72
	Referencias.....	75
	Anexo A.....	80
	Anexo B	81
	Anexo C.....	88
	Anexo D.....	90
	Anexo E	94
	Anexo F	95
	Anexo G.....	120
	Anexo H.....	137

LISTA DE FIGURAS

Figura 1.1 Análisis de las tasas de evolución de ortogrupos WGD	18
Figura 1.2 Esquemas de las tres pruebas de hipótesis probadas	21
Figura 1.3 La mayoría de las proteínas cambian su tasa de evolución después de la WGD	23
Figura 1.4 La mayoría de las proteínas sujetas a una fuerte selección purificadora fuerte aumentan su tasa de evolución después de la WGD	24
Figura 1.5 Las diferentes clases funcionales determinan en gran medida el cambio en la tasa de evolución después de la WGD	25
Figura 1.6 Las diferentes clases funcionales muestran patrones diferentes de evolución después de la WGD	26
Figura 1.7 El desorden intrínseco, la abundancia de proteína y la contribución a la adecuación afectan la tasa de evolución después de la WGD pero son características de la clase funcional	28
Figura 1.8 Hay una aceleración inmediata en las tasas de evolución después de la WGD seguida por una estabilización	31
Figura 1.9 La mitad de los ortogrupos presentan tasas de evolución asimétricas después de la WGD	33
Figura 1.10 La asimetría en las tasas de evolución de los parálogos WGD no está relacionada con expresión diferencial	34
Figura 2.1 Estrategia experimental para intercambiar regiones entre parálogos	44
Figura 2.2 Calibración de las diferentes condiciones usadas en la competencia	46
Figura 2.3 Replicabilidad técnica del ensayo de competencia para determinar adecuación relativa	49
Figura 2.4 Comparación de la adecuación de las mutantes de pérdida de función de los parálogos WGD	50
Figura 2.5 Los diferentes escenarios evolutivos de divergencia funcional de parálogos después de la duplicación	53
Figura 2.6 Mediante una estrategia de complementación genética cuantificamos un índice de divergencia de cada región génica	54
Figura 2.7 El impacto fenotípico de la divergencia codificante y <i>cis</i> -regulatoria después de la duplicación génica	55
Figura 2.8 Cepas descartadas por diversas razones	56
Figura 2.9 Comparación directa de la divergencia fenotípica de la región codificante y región <i>cis</i> -regulatoria	57

Figura 2.10 Resultado de los ensayos de viabilidad de la complementación genética de parejas parálogas con divergencia funcional extrema	59
Figura 2.11 Comparación directa de la divergencia fenotípica de la región codificante y región <i>cis</i> -regulatoria de parejas parálogos con divergencia funcional extrema	61
Figura 2.12 La relación entre dosis génica y adecuación de los parálogos estudiados	63

RESUMEN

Existe evidencia extensiva de que las duplicaciones de genoma completo (WGD) han modelado los genomas de plantas, vertebrados y hongos. Se ha sugerido que del traslape funcional entre parálogos genera robustez y con ello se relaja restricciones sobre las secuencias permitiendo que las proteínas diverjan de su función ancestral y exploren nuevos terrenos funcionales. En el presente trabajo se investigaron los factores que afectan la presión de selección sobre las secuencias codificantes, así como el efecto fenotípico de la divergencia funcional de genes duplicados por WGD.

En la primera parte de este trabajo realicé comparación de las tasas de evolución de 462 parálogos antes y después de la WGD en levaduras hemiascomicetas. La mayoría de las proteínas experimentan una relajación significativa en la presión de selección inmediatamente después de la duplicación; lo cual es especialmente evidente para genes sometidos a una fuerte selección purificadora antes de la WGD. El cambio en las tasas de evolución después de la duplicación aparentemente está asociado con ciertas propiedades de las proteínas como abundancia, contenido de desorden intrínseco y contribución a la adecuación. Sin embargo, cuando si dividimos nuestro grupo de proteínas en diferentes categorías funcionales (metabolismo, ribosomas y regulación) esta asociación desaparece, lo que sugiere que dicha correlación es producto del enriquecimiento de las diferentes clases funcionales y no directamente de la tasa de evolución. En conjunto, estas observaciones sugieren que la categoría funcional, así como la presión de selección previa a la duplicación, son los determinantes más importantes del cambio en tasa de evolución después de la duplicación y muy posiblemente de la divergencia funcional producto de tales cambios.

En la segunda parte de este trabajo se estimó la contribución individual de los cambios en la secuencia codificante y en la región *cis*-regulatoria a la divergencia funcional de un grupo de 20 parejas de genes parálogos WGD. Con esta finalidad, se diseñó una estrategia basada en complementación genética para calcular un índice de divergencia funcional para cada región basado en datos experimentales de alta resolución. Dicho índice indica el efecto directo en la adecuación de la divergencia de cada región, y puede ser evaluado en diferentes condiciones.

Los resultados indican que la divergencia funcional está ampliamente extendida en ambas regiones génicas y es altamente dependiente de la condición de crecimiento. Como es de esperarse, la divergencia funcional de proteína y promotor están significativamente correlacionadas. Sin embargo, los datos obtenidos indican que las proteínas parálogas presentan un índice de divergencia funcional mayor que las regiones *cis*-regulatorias. Todo esto sugiere que, a pesar de que la regulación génica diverge rápidamente después de la duplicación, nuestros resultados sugieren que los cambios más discretos en la región codificante dirigen la divergencia funcional de los genes analizados.

ABSTRACT

There is evidence that Whole Genome Duplications (WGD) has shaped the genomes of plants, vertebrates and fungi. It has been suggested that the functional overlap between paralogs releases selection pressures over protein sequences, allowing them to diverge from the ancestral function and explore new functional paths. In the present work, I proposed to investigate the factors that affect such selection pressure on the coding sequences after the duplication and the nature of the phenotypic effects of such the functional divergence genes duplicated by WGD.

In the first chapter, I made a comparison of the evolutionary rates of 462 WGD-paralogs before and after the WGD of hemiascomycete yeasts. Most proteins underwent a significant relaxation in the selection pressure immediately after the duplication; which was especially evident for genes under a strong purifying selection. The change in evolutionary rates after the duplication is apparently associated with certain properties of the protein like such as abundance, intrinsic disorder content and fitness contribution. Nevertheless, when the data set was split into different functional categories (metabolism, ribosomes and regulation), this associations disappear, suggesting that such a correlation is a by-product of the over or under-representation of a functional category and not directly of due to the evolutionary rate. These observations suggest that the functional category and the selection pressure before the duplication are the main determinants of the fate of paralogs after the WGD.

In the second chapter of this work, I estimate the individual contribution of the changes in coding and *cis*-regulatory sequences to the functional divergence of a group of 20 WGD-paralogous pairs. To assess this question, I designed a strategy based on genetic complementation to estimate a functional divergence index for each region based on high resolution experimental data. The functional divergence index indicates the direct fitness effect of each genetic region, evaluated under multiple environmental conditions. The results revealed significant functional divergence in both genetic regions, and highly dependent on the environment. As expected, the functional divergence of proteins and promoter regions are significantly correlated. Nevertheless, our data show that the paralogous proteins typically present a larger divergence index than their correspondent corresponding *cis*-regulatory regions.

INTRODUCCIÓN

INNOVACIÓN FUNCIONAL Y DUPLICACIÓN

La gran diversidad de formas y colores que observamos en la naturaleza es producto de un proceso de innovación constante a nivel fenotípico. Lo que nos lleva a una de las preguntas centrales en la biología evolutiva: ¿cómo se generan y establecen nuevas funciones y con ello nuevos fenotipos? En este sentido, la duplicación génica es considerada como la fuente primordial de generación de nuevas funciones, y por tanto de nuevos fenotipos.

Con la secuenciación masiva de genomas en las pasadas dos décadas se hizo claro que la duplicación génica es ubicua en prácticamente todas las especies, en lo que denota su importancia como fuente de innovación genética. El 28% de los genes de un genoma promedio se encuentran en dos o más copias. La relevancia evolutiva de este evento reside en que la aparición de una copia extra puede facilitar la adquisición de funciones novedosas sin perder la función original (Jianzhi Zhang, 2003). Lynch y Conery estimaron que la duplicación de un gen ocurre (y se fija dentro de las poblaciones) aproximadamente una vez por cada gen en 100 millones de años, en eucariontes como *Homo sapiens*, *Mus musculus*, *Drosophila melanogaster*, *Caenorhabditis elegans*, *Arabidopsis thaliana* y *S. cerevisiae* (Lynch & Conery, 2000).

Según el modelo clásico, la mayoría de los genes están sujetos a múltiples presiones de selección, la duplicación génica permite que en una o las dos copias del gen se liberen una parte de la presión selectiva y puedan explorar nuevos terrenos funcionales (Hahn, 2009; Ohno, 1970). Los genes duplicados atraviesan por diversas etapas que determinan su destino funcional y por lo tanto su preservación a lo largo de tiempos evolutivos prolongados. La

historia natural de un gen duplicado puede dividirse en tres etapas principales: duplicación, fijación en la población y preservación (Innan & Kondrashov, 2010).

LA DUPLICACIÓN DE GENES SUCEDE EN DIFERENTES ESCALAS

Duplicaciones a pequeña escala

La duplicación a pequeña escala (SSD, por sus siglas en inglés), es usualmente producto de la duplicación de segmentos pequeños de cromosoma que dan lugar a la duplicación de uno o más genes. Entre los mecanismos de duplicación en pequeña escala destacan la *retrotransposición*, el *sobrecruzamiento desigual* y la *duplicación de segmentos no homóloga*. La *retrotransposición* se debe a que la “retrotranscriptasa” convierte a un mRNA en un cDNA y éste fragmento de DNA se inserta en el genoma, debido a que se requiere de la aparición de un promotor para que esta copia sea funcional éstas secuencias rara vez son activas. Tanto el *sobrecruzamiento desigual* como la *duplicación no homóloga de segmentos* son producto de errores durante la recombinación homóloga y en conjunto son la principal fuente de generación de genes duplicados en pequeña escala (Conant & Wolfe, 2008).

Entre los grupos funcionales que están enriquecidos en el grupo de genes duplicados por SSD se encuentran: los genes que codifican para transportadores de azúcares así como aquellos con actividades de helicasa e hidrolasa; también genes de proteínas de unión como las involucradas unión a ácidos nucleicos y regulación transcripcional (Hakes, Pinney, Lovell, Oliver, & Robertson, 2007).

La duplicación de genoma completo

La duplicación de genoma completo (WGD, por sus siglas en inglés), también llamada poliploidización, se refiere a la adquisición un set adicional de cromosomas (Otto, 2007). La ganancia del juego extra de cromosomas puede ser producto de un error llamado *no disyunción*; éste sucede cuando los cromosomas homólogos fallan en separarse en la meiosis o durante la mitosis si las cromátidas hermanas no se separan. Cuando ambos juegos de cromosomas provienen de individuos de la misma especie se denomina “autopoliploide”, mientras que cuando cada set de cromosomas proviene de una especie diferente (hibridización) se denomina “alopoliploide”.

Los eventos de duplicación de genoma completo han sido determinantes en la historia evolutiva de los organismos eucariontes y las pruebas de ello se pueden observar en los linajes más estudiados. En plantas se han reportado múltiples rondas de duplicación de genoma, en particular, una de ellas ha sido relacionada con surgimiento y radiación de angiospermas (Adams & Wendel, 2005). En vertebrados hay evidencias de al menos dos rondas de duplicación de genoma (Dehal & Boore, 2005). Por último, en hongos ascomicetos hay reportada una ronda de duplicación de genoma en el linaje *S. cerevisiae* (Kellis, Birren, & Lander, 2004; Wolfe & Shields, 1997).

Los genes que se conservan después de la duplicación del genoma pertenecen a grupos funcionales radicalmente diferentes a los de las duplicaciones sencillas. Éstos están enriquecidos en genes de metabolismo central como glicólisis y síntesis de aminoácidos, genes que codifican para proteínas ribosomales y genes pertenecientes a complejos proteicos, entre otros (Guan, Dunham, & Troyanskaya, 2007). Además, los parálogos WGD tienden a generar un efecto fenotípico menor que los duplicados SSD cuando se pierden, posiblemente

debido a que conservan mayor traslape funcional (Hakes et al., 2007). Debido a dicho traslape funcional los genes WGD comparten más interacciones físicas que genes SSD. Probablemente la diferencia más importante es que los duplicados WGD son más propensos a pertenecer a complejos proteínicos que los duplicados SSD. Además los genes WGD que pertenecen a complejos son menos propensos a ser esenciales que los genes SSD que pertenecen a este mismo grupo de genes (Hakes et al., 2007). De hecho, se ha sugerido que dicha diferencia en contenido de genes pertenecientes a complejos es responsable en gran parte de las diferencias de los grupos funcionales retenidos por cada tipo de duplicación.

CIERTOS TIPOS DE GENES SE DUPLICAN MÁS FRECUENTEMENTE

Existen propiedades moleculares que pueden afectar la frecuencia en que los genes se duplican y se fijan en las poblaciones:

Los genes menos importantes se fijan por duplicado más frecuentemente. Un análisis comparativo entre un grupo de genes “singletons” (genes de copia única) y un grupo de genes duplicados en *S. cerevisiae* y cuatro especies cercanas reveló que los genes más dispensables tienden a duplicarse más que sus contrapartes. Mientras que el grupo de singletons está enriquecido en genes esenciales (28%) y la pérdida de estos genes provoca una reducción en la adecuación del 38% en promedio; el grupo de duplicados presenta un porcentaje menor de genes esenciales (17%) y una reducción en la adecuación de solo 24% (Li, Huang, Xia, & Sun, 2006).

La conectividad en las redes de interacción de proteínas. Se sabe que la duplicación génica juega un papel muy importante en el crecimiento de las redes de interacción proteína-proteína. Sin embargo, se ha demostrado que los genes que tienen una alta conectividad en redes de

interacción, como los genes de procesamiento de RNA y DNA que se encuentran en el núcleo, tienden a duplicarse menos con respecto a otros grupos funcionales, mientras que genes que tienen una conectividad menor, como las proteínas transportadoras de la membrana, poseen tienden a duplicarse más frecuentemente (Li et al., 2006). Estos resultados sugieren que la alta conectividad de ciertos genes restringe de manera inmediata su capacidad de ser fijados en la población en más de una copia.

El balance de dosis y la duplicabilidad. La teoría del balance de dosis propone que los desbalances estequiométricos en complejos macromoleculares pueden provocar efectos fenotípicos que son dependientes de la dosis génica de las subunidades proteicas que componen dicho complejo (Reiner A Veitia, Bottani, & Birchler, 2008). Estos fenotipos pueden ser producto tanto de la disminución en la expresión como de la sobreexpresión de las subunidades de complejos proteicos. Una evidencia importante de este fenómeno es que las subunidades de complejos importantes son a menudo coexpresadas o presentan perfiles de expresión altamente correlacionados, por lo que es de esperar que la dosis de dichos genes sea altamente regulada. De hecho, cuando suceden perturbaciones de este equilibrio en la dosis génica, como las aneuploidias o la duplicación génica, a menudo se observan defectos en el crecimiento (Reiner A Veitia et al., 2008; R. A. Veitia & Potier, 2015). La escala de la duplicación génica influye en el efecto de balance de dosis, por ejemplo, en las duplicaciones de genoma completo se ha observado una retención más elevada (con respecto a duplicados SSD) de parálogos que pertenecen a complejos. Al duplicarse todas las subunidades en un solo evento, la estequiometría no se ve afectada y como consecuencia no hay un efecto deletéreo por el desbalance de dosis. Por otro lado, no se observa este mismo comportamiento en duplicaciones de segmentos pequeños para los que, por el contrario, hay una

representación mucho menor de estos grupos de genes entre los genes duplicados (Maere et al., 2005).

LOS DESTINOS POSIBLES DE UN GEN RECIÉN DUPLICADO

Inmediatamente después de la duplicación ambas copias del gen son funcionalmente redundantes, lo cual implica que la inactivación de uno de estos genes tendrá poco o nulo efecto sobre el fenotipo. La falta de restricciones selectivas provoca que los duplicados sean presas fáciles ante las mutaciones degenerativas que provocan su pérdida de función, de hecho, este es el destino de la gran mayoría de las copias de genes. Los genes duplicados pueden tener diferentes destinos funcionales, entre los cuales destacan los siguientes:

Redundancia funcional. Algunos genes se retienen por un efecto positivo ante el aumento de la dosis génica; una cantidad más alta del producto de tales genes genera ventaja adaptativa para el organismo. En este caso ambas copias estarán protegidas de las mutaciones degenerativas y típicamente exhiben una tasa de evolución baja (Krakauer & Nowak, 1999; Nowak, Boerlijst, Cooke, & Smith, 1997). Otra alternativa es que la copia extra actúe como respaldo de la función ancestral ante mutaciones deletéreas proporcionando robustez a la función y al sistema genético (Innan & Kondrashov, 2010).

Subfuncionalización. A menos que una cantidad extra de producto resulte ventajosa, la presencia de dos genes con funciones idénticas tiene poca estabilidad en el genoma ya que hay una relajación en la presión de selección. No obstante, ambas copias se pueden mantener estables cuando difieren en algunos aspectos de su función (Nowak et al., 1997). La subfuncionalización es un proceso en el que ambos duplicados se reparten funciones bioquímicas, perfiles de expresión y/o interacciones físicas. El modelo más aceptado de

subfuncionalización se denomina Duplicación Degeneración Complementación, las dos copias acumulan mutaciones neutrales, pero degenerativas, a tal punto que cada una pierde parte de la función original y es incapaz de desempeñar esta función completamente. Ambas copias pueden ser preservadas ya que se han dividido la carga funcional del gen ancestral y muy posiblemente están sometidas a una selección purificadora (Force et al., 1999).

Neofuncionalización. Se refiere a la adquisición de una función nueva tomando como base la función ancestral del duplicado. En el modelo clásico de neofuncionalización, hay una pérdida de la funcionalidad original de alguna de las dos copias debido a mutaciones, en un proceso llamado *pseudogenización* (A. L. Hughes, 1994); mientras que durante este periodo, también suceden mutaciones que confieren una nueva función a esta copia no funcional, este modelo se denomina Mutación Durante la No-funcionalidad (MND) (Innan & Kondrashov, 2010).

Pseudogenización. Finalmente, el destino más común de un gen recién duplicado es la pérdida de función debido a la acumulación de mutaciones que provocan pérdida de función, producto de una relajación en la presión de selección que es provocada por la redundancia funcional entre ambas copias.

NUESTRO MODELO DE INVESTIGACIÓN DE DUPLICACIÓN DE GENOMA COMPLETO: *Saccharomyces cerevisiae*

La levadura *S. cerevisiae* ha sido utilizada por milenios para la producción de bebidas alcohólicas, incluyendo cerveza, vino y destilados, o para la elaboración de pan (Cavalieri, McGovern, Hartl, Mortimer, & Polsinelli, 2003; Samuel, 1996), contribuyendo de manera importante al desarrollo de la civilización como la conocemos. Además de la industria

alimenticia, *S. cerevisiae* es uno de los modelos eucariontes más estudiados por lo que tiene el privilegio de haber sido el primer eucarionte en ser secuenciado por completo en 1996 (Goffeau et al., 1996).

El genoma de *S. cerevisiae* contiene 12 Mb repartidas en 16 cromosomas que codifican para cerca de 6000 genes (Goffeau et al., 1996). Solo 20% de esos genes son esenciales para el crecimiento en condiciones estándar de laboratorio (Giaever et al., 2002). El 30% de los genes están presentes en más de una copia (Goffeau et al., 1996; Wolfe & Shields, 1997), de los cuales aproximadamente la mitad provienen de un evento de duplicación de genoma completo. Se calcula que esta WGD sucedió hace unos 100 millones de años y resultó en la duplicación de aproximadamente 10% del genoma en el caso de *S. cerevisiae* (Kellis et al., 2004; Wolfe & Shields, 1997).

S. cerevisiae puede sobrevivir y dividirse de manera estable tanto en estado haploide como diploide. Se reproduce la mayor parte del tiempo de manera asexual por mitosis, pero cuando hay escasez de nutrientes (nitrógeno y fuente de carbono), las levaduras diploides realizan meiosis generando esporas haploides (Herskowitz, 1988). Se ha observado que en la naturaleza las poblaciones de levadura son mayormente clonales, se calcula que hay cruza exogámicas en muy baja frecuencia (una vez cada 50,000 divisiones) (Ruderfer, Pratt, Seidel, & Kruglyak, 2006). Las ventajas que proporciona dicho estilo de vida de *S. cerevisiae* han permitido la sistematización de múltiples técnicas de biología molecular a gran escala; como generación de mutantes dirigida por recombinación homóloga; y generación de mutantes dobles mediante la cruce de mutantes sencillas. Como resultado, existen múltiples recursos genómicos disponibles: colección de todas las mutantes sencillas de pérdida de función (Giaever et al., 2002); colección de fusiones transcripcionales individuales con un marcador

fluorescente (Ghaemmaghami et al., 2003); múltiples librerías de plásmidos que contienen genes individuales expresados bajo el promotor endógeno (Ho et al., 2009), entre otros. Tales recursos permiten un enfoque de sistemas al estudio de problemas de ciencia básica y aplicada; haciendo de *S. cerevisiae* una herramienta sin igual para la biología evolutiva.

OBJETIVOS

OBJETIVO GENERAL

Estudiar la dinámica evolutiva de las secuencias codificantes y la divergencia funcional regulatoria y codificante de genes duplicados originados por la duplicación de genoma completo (WGD) en la levadura *S. cerevisiae*.

OBJETIVOS ESPECÍFICOS

-) Describir los patrones de cambio en la presión de selección sobre la secuencia codificante, así como las propiedades moleculares y fenotípicas relacionadas con cambios en la presión de selección en parálogos después de la WGD.
-) Evaluar la contribución relativa a la adecuación de las regiones codificante y *cis*-regulatoria a la divergencia fenotípica de genes duplicados mediante una aproximación experimental.

CAPÍTULO 1

ESTUDIO DE LA DINÁMICA EVOLUTIVA DE GENES PARÁLOGOS DESPUÉS DE UNA DUPLICACIÓN DE GENOMA COMPLETO¹

Es generalmente aceptado que la mayor parte de las parejas de genes parálogos aceleran su tasa de evolución después de la duplicación debido al traslape funcional que libera parte de la presión de selección que actúa sobre sus secuencias. Sin embargo, es claro que no todos los parálogos se comportan igual y pocos trabajos se han dado la tarea de describir las propiedades génicas que explican las diferencias en la forma en que los parálogos evolucionan después de la duplicación. Incluso genes que provienen del mismo evento de duplicación, como es el caso de los genes que provienen de la WGD en levadura, presentan una multitud de comportamientos diferentes.

1.1 ANTECEDENTES

Hay ciertas características funcionales asociadas con genes que se retienen en dos copias después de una WGD. Por ejemplo, en *S. cerevisiae* se favoreció la retención de proteínas ribosomales, enzimas de metabolismo de fuentes de carbono y cinasas de transducción de señales (Seoighe & Wolfe, 1999). De manera interesante, las categorías funcionales de los genes retenidos en duplicado después de la WGD son similares a lo largo de diferentes grupos taxonómicos desde levaduras hasta *Arabidopsis thaliana* o *Paramecium spp.* (Conant &

¹ Este capítulo contiene texto y figuras tomados o modificados del trabajo "Increased rates of protein evolution and asymmetric deceleration after the whole-genome duplication in yeasts" (Ascencio et al., en revisión; Anexo F)

Wolfe, 2008). Los parálogos WGD son menos esenciales que los genes codificados en una sola copia (singletons) debido a que retienen un traslape funcional importante (Dean, Davis, Davis, & Petrov, 2008; DeLuna et al., 2008). Sin embargo, numerosos casos de tasas de evolución asimétricas y pérdida de la capacidad de compensación sugieren que también existe neo-funcionalización extensiva después de la WGD (Byrne & Wolfe, 2007).

Varios estudios han tratado de describir el efecto de la duplicación sobre la selección natural. Dichos trabajos han sugerido que las tasas de evolución, de las secuencias parálogas, tienden a aumentar después de la duplicación (Davis & Petrov, 2004; Kondrashov, Rogozin, Wolf, & Koonin, 2002; Lynch & Conery, 2000; Scannell & Wolfe, 2008). Sin embargo, no se conocen los factores que determinan el cambio en la presión de selección que actúa sobre las secuencias codificantes. Por otro lado, se ha propuesto que el nivel de expresión, relevancia funcional y densidad funcional son los factores determinantes de la tasa de evolución de las proteínas en general. Por lo que, genes que se expresan en altos niveles, son esenciales para el crecimiento, y/o que tienen alta densidad funcional tienden a evolucionar a tasas más bajas que el resto (Choi & Hannenhalli, 2013; Gout, Kahn, Duret, & Consortium, 2010; J. Zhang & Yang, 2015). Sin embargo, no está claro si dicho patrón se mantiene para las diferentes clases funcionales de los genes que se retienen en dos copias por largos periodos evolutivos para el caso particular de WGD.

En este capítulo, se investigó el cambio en las tasas de evolución en a diferentes puntos de la historia evolutiva de un set de parejas de genes parálogos, prestando especial atención a las ramas antes y después de la WGD. También se determinó la relación de las funciones bioquímicas, características moleculares o fenotípicas de las proteínas con patrones de evolución similares después de la duplicación.

1.2 HIPÓTESIS

Existe evidencia extensiva de que hay un cambio en las tasas de evolución después de la duplicación, sin embargo, los factores que influyen en este cambio son desconocidos. Se ha reportado que la presión de selección que actúa sobre las secuencias codificantes es en gran parte determinada por las propiedades moleculares de las proteínas como abundancia, estructura tridimensional y contribución a la adecuación. Es probable que la relajación en la tasa de evolución seguida de la duplicación determinada en gran parte por dichas características generales.

1.3 MÉTODOS

1.3.1 Selección de ortogrupos, secuencias, árboles filogenéticos y alineamientos usados

Usamos la base de datos Fungal Orthogroups Repository (<http://www.broadinstitute.org/regev/WGDOs/>), que contiene secuencias de 23 especies de ascomicetos (Ilan Wapinski, Avi Pfeffer, Nir Friedman, & Aviv Regev, 2007). Ésta es una base de datos automatizada generada usando el algoritmo SYNERGY (I. Wapinski, A. Pfeffer, N. Friedman, & A. Regev, 2007), el cual usa similitudes en las secuencias y la filogenia de las especies para reconstruir la historia evolutiva de una familia de genes. Como resultado, se obtienen árboles filogenéticos que reconstruyen fielmente la trayectoria evolutiva de los todos los genes de estas especies, incluyendo duplicaciones y pérdidas en los linajes estudiados. Cada filogenia contiene todos los ortólogos presentes en las especies de la base de datos, los cuales son denominados como *ortogrupos* (Ilan Wapinski et al., 2007).

Para este trabajo se tomaron 545 ortogrupos que contienen genes que se retuvieron en duplicado en *S. cerevisiae* en conjunto con sus ortólogos pre y post-WGD. De aquí en adelante me refiero a estos grupos de genes como ortogrupos-WGD. En la **Figura 1.1** mostramos el ortogrupo de *PSK1/PSK2* de *S. cerevisiae* como ejemplo.

Descartamos 83 ortogrupos debido a la baja calidad (alta frecuencia de nucleótidos indeterminados y *gaps*) y/o cantidad (menos de seis secuencias de ortólogos) de las secuencias. Tomamos los 462 ortogrupos-WGD restantes para realizar análisis posteriores. Nos aseguramos que los marcos de lectura estuvieran asignados correctamente comparando las secuencias de proteínas y nucleótidos usando PAL2NAL (Suyama, Torrents, & Bork, 2006). Finalmente, alineamos las secuencias de cada ortogrupo con MUSCLE 3.6 (Edgar, 2004) usando parámetros estándar.

Ortogrupo: *PSK1/PSK2*

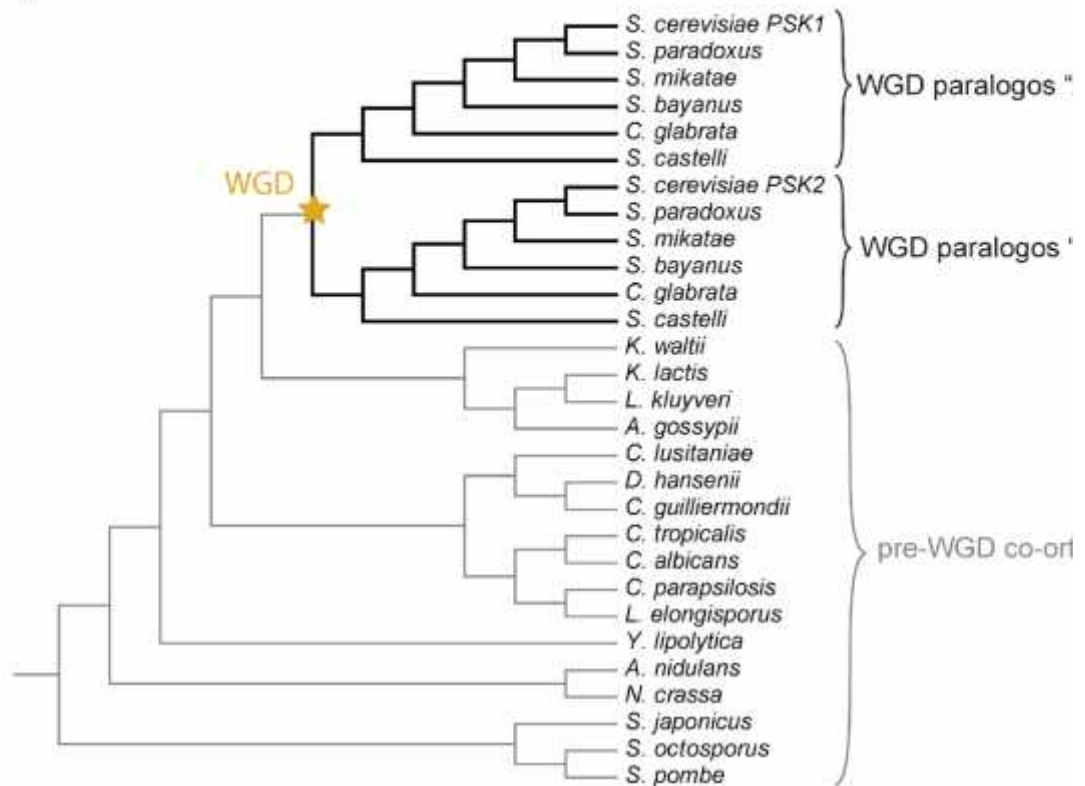


Figura 1.1 Análisis de las tasas de evolución de ortogrupos WGD. Mostramos un ejemplo de un árbol filogenético de un ortogrupo de los parálogos *PSK1/PSK2* de *S. cerevisiae*. Las ramas grises indican los ortólogos de *PSK1/PSK2* en especies de ascomicetos que divergieron antes de la WGD (estrella amarilla), mientras que las ramas negras representan las especies post-WGD con dos clados parálogos.

1.3.2 Pruebas filogenéticas con máxima verosimilitud con CodeML

Con los 462 alineamientos y sus correspondientes árboles filogenéticos, probamos tres diferentes modelos evolutivos usando CodeML, un paquete de PAML 4.4 (Yang, 2007). Para probar diferencias antes y después la duplicación, en la primera *primera* hipótesis se identificó el nodo de la WGD con base en la topología de cada árbol filogenético;

posteriormente, agrupamos las ramas en pre y post-WGD **Figura 1.2A**. Con esta topología, el alineamiento y el árbol filogenético correspondiente corrimos CodeML usando un archivo control estándar (ver ejemplo **Anexo A**). Para la hipótesis alternativa, usamos el modelo de ramas especificado por el usuario, el cual permite calcular tasas de evolución (dN/dS) diferentes para las ramas pre y post-WGD. Éste hipótesis alternativa fue comparado con una hipótesis nula: todas las ramas tienen la misma tasa dN/dS (). CodeML proporciona los valores de verosimilitud para cada prueba de hipótesis así como las respectivas tasas . Usamos una prueba de razón de verosimilitud (LRT) para comparar la hipótesis alternativa con la hipótesis nula de la siguiente manera:

$$L = 2 * (L_1 - L_0)$$

donde:

L_1 es valor de verosimilitud de la hipótesis alternativa

L_0 es valor de verosimilitud de la hipótesis nula

Una vez obtenido el LRT se usa para calcular el p -value usando una distribución χ^2 con un grado de libertad; la hipótesis nula es rechazada con un corte de $p < 0.05$.

Para buscar diferencias en las ramas post-duplicación realizamos una **segunda** prueba de hipótesis. En la hipótesis alternativa consideramos diferentes tasas para tres grupos de ramas: las ramas pre-WGD (0), las ramas que siguen inmediatamente después de la WGD (1) y otra para los dos clados modernos (2), ver **Figura 1.2-B**. En este caso la hipótesis nula: consideramos solo dos tasas una pre-WGD y otra post-WGD (1 = 2). Por último, para detectar diferencias de tasas en los clados parálogos modernos realizamos una **tercera** prueba de hipótesis; en la hipótesis alternativa consideramos diferentes tasas cuatro

diferentes grupos de ramas: las ramas pre-WGD (0), las ramas que siguen inmediatamente después de la WGD (1) y cada uno de los clados parálogos modernos (3 y 4), ver **Figura 1.2-C**. Por el contrario, en la hipótesis nula consideramos tasas iguales para los dos clados parálogos (3 = 4). El LRT y el *p-value* fueron calculados como en el caso anterior.

1.3.3 Grupos funcionales, abundancia, adecuación y desorden intrínseco

Las categorías funcionales fueron definidas manualmente basándonos en las anotaciones funcionales para *S. cerevisiae* en la *Saccharomyces Genome Database* (<http://www.yeastgenome.org/>). La categoría *regulatoria* incluye genes anotados como cinasas y factores de transcripción. *Metabolismo* incluye genes que codifican para enzimas involucradas en procesos biosintéticos y catabólicos. La categoría *ribosomal* incluye genes que codifican para componentes estructurales del ribosoma.

Usamos los datos de abundancia de proteínas de *S. cerevisiae* determinados mediante microscopía de fluorescencia (Breker, Gymrek, & Schuldiner, 2013). Los datos de adecuación fueron tomados de un trabajo en el que se calcularon mediante un método de alta resolución y relativa a una cepa silvestre (Breslow et al., 2008). Finalmente, los datos de contenido de desorden intrínseco de proteínas fueron tomados del trabajo de Landry et al. (Landry, Levy, & Michnick, 2009), donde las regiones intrínsecamente desordenadas se predijeron usando el algoritmo DISOPRED (Jones & Ward, 2003).

Todas las tasas dN/dS, los valores LRT y propiedades de las proteínas pueden ser encontradas en la tabla del **Anexo B**.

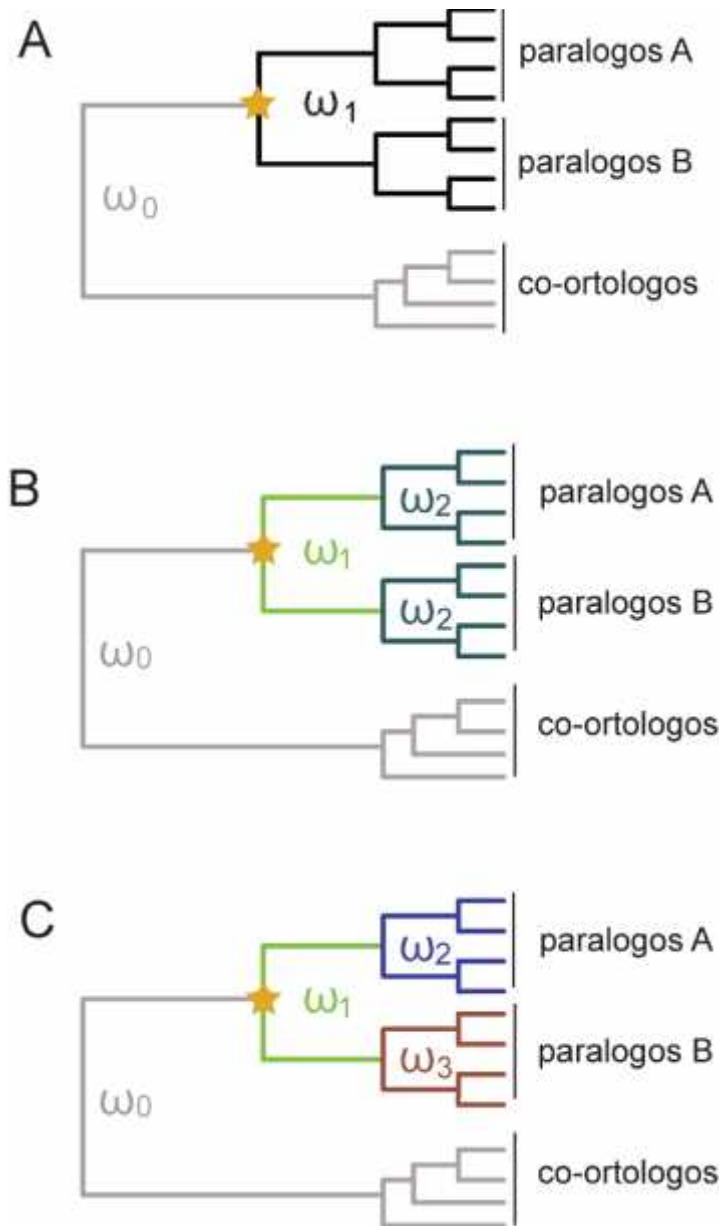


Figura 1.2 Esquemas de las tres pruebas de hipótesis probadas. Cada esquema representa las filogenias de los ortogrupos WGD. La estrella amarilla representa la WGD. Los diferentes colores corresponden a los grupos de ramas cuyas tasas de evolución se comparan en las tres pruebas de hipótesis. **(A)** La hipótesis alternativa de la primera prueba de hipótesis; dos tasas de evolución: una pre-WGD (ω_0 en gris) y otra en las ramas post-WGD (ω_1 en negro). **(B)** La hipótesis alternativa de la segunda prueba de hipótesis; tres tasas de evolución: una pre-WGD (ω_0 en gris), otra inmediatamente post-WGD (ω_1 en verde claro) y una última para las ramas subsecuentes (ω_2 en azul-verde). **(C)** La hipótesis alternativa de la tercera prueba de hipótesis; cuatro tasas de evolución: una pre-WGD (ω_0 en gris), otra inmediatamente post-WGD (ω_1 en verde claro) y una última para cada clado parálogo moderno (ω_2 y ω_3 en azul y en rojo respectivamente).

1.4 RESULTADOS

1.4.1 La mayor parte de los ortogrupos experimentan un cambio en la tasa de evolución después de la WGDs

El modelo propuesto en los 70's por Ohno sugiere que los genes recién duplicados, al ser capaces de compensarse funcionalmente, experimentan un incremento en sus tasas de evolución después de su duplicación (A. L. Hughes, 1994; Ohno, 1970). Nos preguntamos hasta qué punto dicho modelo se aplica en el caso de los linajes que descienden de la WGD en hemiascomicetos. Con esta finalidad, usamos un método de *máxima verosimilitud* para probar una hipótesis alternativa de tasas dN/dS diferentes para los linajes pre y post-WGD; comparándola con la hipótesis nula en la que las tasas son iguales. Realizamos dichas pruebas para 426 ortogrupos para los cuales se retuvieron dos parálogos después de la WGD en *S. cerevisiae*.

Encontramos que las tasas de evolución de las ramas pre-WGD y post-WGD son significativamente diferentes para la mayoría (>75%) de los ortogrupos analizados (**Figura 1.3-A**, barras naranjas). Mientras que los ortogrupos restantes presentaron un cambio no significativo en sus tasas de evolución después de la WGD por lo que la hipótesis nula no pudo ser rechazada (**Figura 1.3-A**, barras azules). Las tasas dN/dS promedio son generalmente menores a 1 pero casi siempre mayores para las ramas post-WGD ($p < 10^{-40}$, prueba de rangos de Wilcoxon), lo que sugiere una relajación significativa en la selección purificadora después de la WGD (**Figura 1.3-B**). Estos datos confirman que la mayoría de los genes duplicados por WGD cambian significativamente su tasa de evolución después de la duplicación.

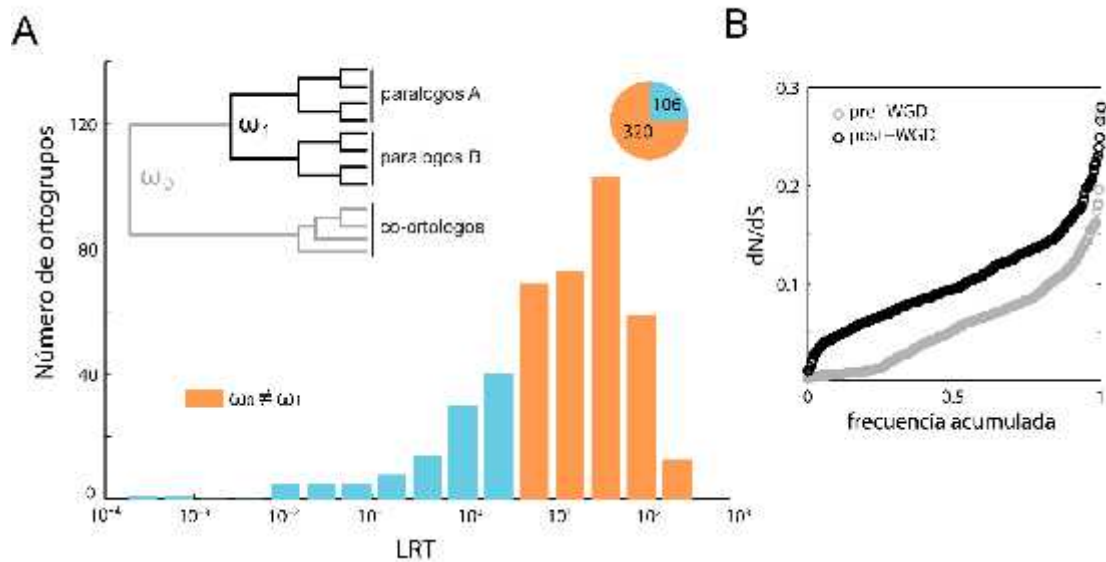


Figura 1.3 La mayoría de las proteínas cambian su tasa de evolución después de la WGD (A) Distribución de frecuencias de los valores de las *pruebas de razón de verosimilitud* (LRT) para la primera prueba de hipótesis. Las barras color naranja indican ortogrupos para los cuales la hipótesis nula, de una sola tasa dN/dS en toda la filogenia, es rechazada ($p < 0.05$); las barras azules indican ortogrupos para los cuales la hipótesis nula no pudo ser rechazada. (B) Graficas de la distribución de frecuencia acumulada de las tasas dN/dS de los linajes pre y post-WGD.

Comparamos las tasas dN/dS de las ramas pre-WGD y post-WGD de cada ortogrupos-WGD estudiado en la **Figura 1.4-A**. Notamos que la gran mayoría de los ortogrupos se sitúan por arriba de la diagonal (las tasas dN/dS pre-WGD son menores que las tasas post-WGD); y solo una fracción se encuentra en el escenario opuesto en el que las tasas de evolución disminuyen después de la duplicación. Si observamos la distribución de las tasas dN/dS de las ramas pre-WGD se pueden distinguir dos sets independientes de ortogrupos; uno sujeto a una presión purificadora muy fuerte ($dN/dS < 0.02$) y otro sujeto a selección purificadora relativamente más débil ($dN/dS > 0.02$), mostramos dichos datos en histogramas amarillos y morados en la **Figura 1.4-A**. Se puede observar que prácticamente todos los ortogrupos sujetos a una selección purificadora intensa previa a la WGD; cambian sus tasas de evolución después de la duplicación (**Figura 1.4-B**).

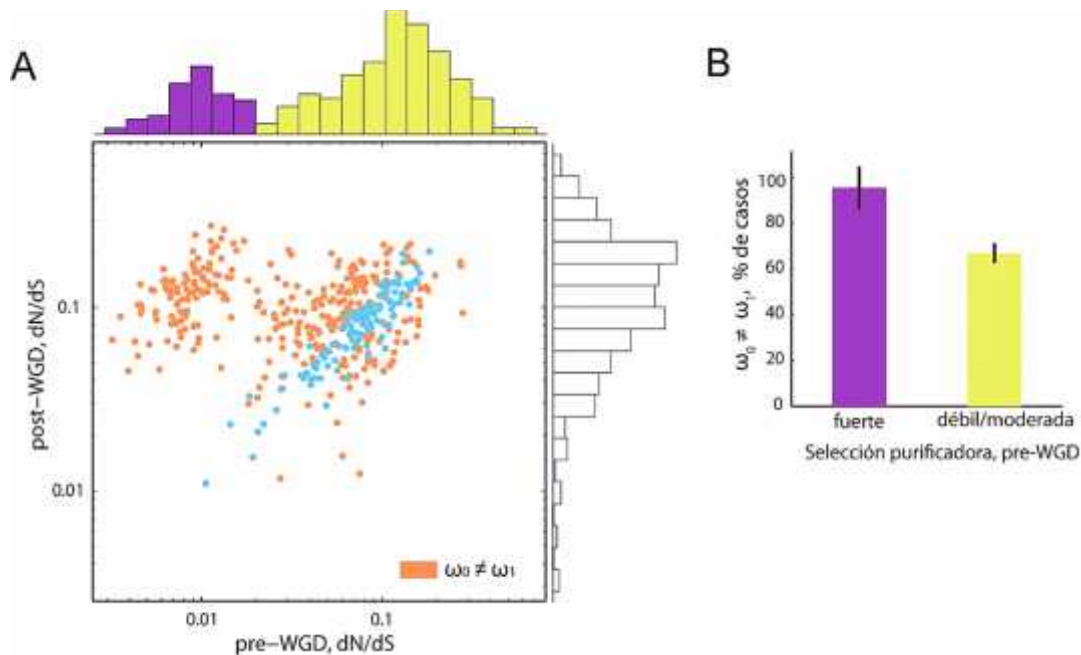


Figura 1.4 La mayoría de las proteínas sujetas a una fuerte selección purificadora fuerte aumentan su tasa de evolución después de la WGD (A) Tasas dN/dS post-WGD comparados con las tasas pre-WGD de 426 ortogrupos. Las barras moradas en el histograma del eje x indican ortogrupos bajo una fuerte selección purificadora previa a la WGD ($dN/dS < 0.02$); mientras que las barras color limón son ortogrupos bajo una débil a moderada selección purificadora pre-WGD. Cada punto representa un ortogrupo individual; los puntos naranjas representan ortogrupos para los cuales la hipótesis nula es rechazada ($\omega_0 \neq \omega_1$), mientras que los puntos azules son ortogrupos para los cuales la hipótesis nula no pudo ser rechazada. (B) Fracción de ortogrupos bajo una selección purificadora fuerte o débil/moderada para los cuales se acepta la hipótesis alternativa ($\omega_0 \neq \omega_1$).

1.4.2 Las proteínas con funciones diferentes muestran tendencias evolutivas diferentes

Como mencionamos anteriormente las proteínas ribosomales, enzimas de metabolismo de carbohidratos, y proteínas regulatorias (especialmente cinasas y factores de transcripción) se retuvieron preferentemente en la WGD en *S. cerevisiae* (Seoighe & Wolfe, 1999). De hecho, el 40% de los ortogrupos estudiados están asociados con alguna de estas tres funciones. Usando estas tres categorías funcionales altamente representadas, se evaluó si las proteínas con diferentes roles bioquímicos presentan patrones evolutivos similares. Los resultados

muestran que la composición de los ortogrupos que cambian su tasa de evolución después de la WGD es determinada en gran parte por las funciones de las proteínas. En particular, los genes regulatorios están sobrerrepresentados en los que cambian su tasa de evolución, mientras que las proteínas ribosomales están empobrecidas en este mismo grupo de genes (**Figura 1.5**).

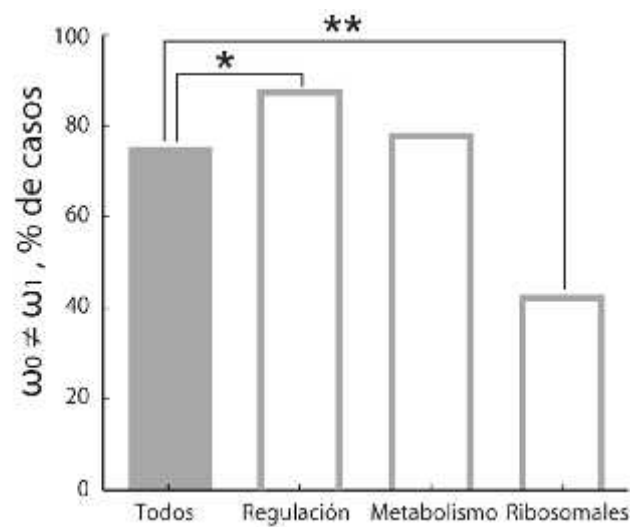


Figura 1.5 Las diferentes clases funcionales determinan en gran medida el cambio en la tasa de evolución después de la WGD. Fracción de ortogrupos para los que la hipótesis alternativa ($\omega_0 \neq \omega_1$) es aceptada; la gráfica muestra todos los ortogrupos ensayados y tres diferentes categorías funcionales. Se muestran las categorías funcionales que están enriquecidas o empobrecidas en el set de ortogrupos que cambian su tasa de evolución, $*p < 0.05$ y $**p < 0.01$ calculados con el test χ^2 .

Las tasas dN/dS incrementan para la gran mayoría de los ortogrupos-WGD y en el caso de las proteínas con función regulatoria, como las cinasas, esta tendencia es incluso exacerbada. En la **Figura 1.6** podemos observar que para la categoría regulatoria la media de las tasas pre-WGD es significativamente diferente a la media de las tasas post-WGD ($p < 10^{-7}$, test de rangos de Wilcoxon). Por el contrario, en el caso de las enzimas de metabolismo y las

proteínas ribosomales la media de las tasas pre y post-WGD no es significativamente diferente ($p=0.22$ and $p=0.61$, respectivamente). De hecho, las tasas de las proteínas regulatorias son ligeramente mayores al resto de los ortogrupos-WGD en las ramas post-WGD, mientras que éste no es el caso de las proteínas regulatorias o enzimas metabólicas, que tienden a evolucionar más lento que otros genes (**Figura 1.6**).

De manera interesante, las proteínas ribosomales evolucionan más rápido que otros genes en las ramas pre-WGD, pero más lentamente después de la duplicación. En conjunto, estos resultados sugieren que la función bioquímica es un factor determinante en los patrones de evolución de la secuencia codificante en genes que retienen en duplicado por mucho tiempo después de la WGD. En particular, la aceleración de las tasas de evolución parece ser específica de ciertas clases funcionales, como las proteínas regulatorias.

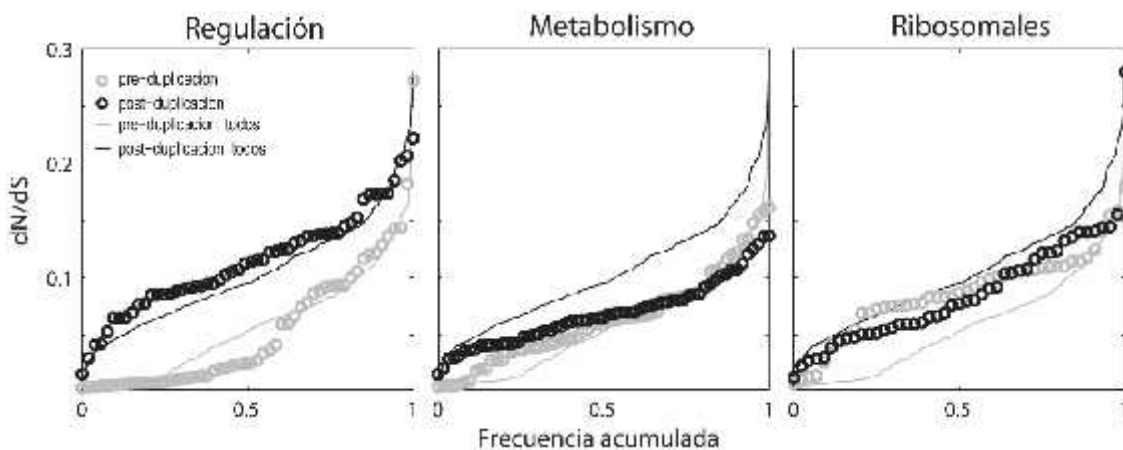


Figura 1.6 Las diferentes clases funcionales muestran patrones diferentes de evolución después de la WGD. Gráfica de distribución acumulada de las tasas dN/dS para tres categorías funcionales. En líneas (todos los ortogrupos) y círculos (categorías específicas) grises las tasas dN/dS pre-WGD; en líneas y círculos negros mostramos las tasas post-WGD.

1.4.3 La abundancia de proteína, el desorden intrínseco y la contribución a la adecuación son factores determinantes en el cambio de las tasas de evolución después de la WGD

La tasa a la que evoluciona una región codificante depende en gran parte del nivel de restricción funcional a la que está sujeta la proteína. En particular, el nivel de expresión es un factor determinante en la tasa de evolución de una proteína, por lo que proteínas muy abundantes evolucionan más lento (Drummond, Bloom, Adami, Wilke, & Arnold, 2005; Pal, Papp, & Hurst, 2001). La selección contra de errores en la traducción, plegamiento de proteínas o interacciones aberrantes son posibles causas de dicha tendencia (J. Zhang & Yang, 2015). Nos preguntamos hasta qué punto la abundancia de proteína explica los cambios en las tasas de evolución después de la duplicación y como se relaciona con la función bioquímica. Con éste objetivo en mente, usamos datos experimentales de abundancia de proteína medidos para *S. cerevisiae* (Breker et al., 2013). Comparamos las distribuciones de abundancias de proteínas entre genes que pertenecen a ortogrupos que cambian y no cambian su tasa de evolución después de la duplicación (**Figura 1.7-A**). En correspondencia con observaciones anteriores, las proteínas con tasas constantes son típicamente más abundantes que las proteínas con tasas diferentes. Como es de esperarse las diferentes categorías funcionales presentan grandes diferencias en abundancia de proteína, por ejemplo, las proteínas ribosomales son mucho más abundantes que las proteínas regulatorias. Sin embargo, la tendencia general que se observa en el set completo de ortogrupos (las proteínas que cambian su tasa después de la duplicación son menos abundantes) no se conserva para el caso de las diferentes categorías funcionales (**Figura 1.7-A**).

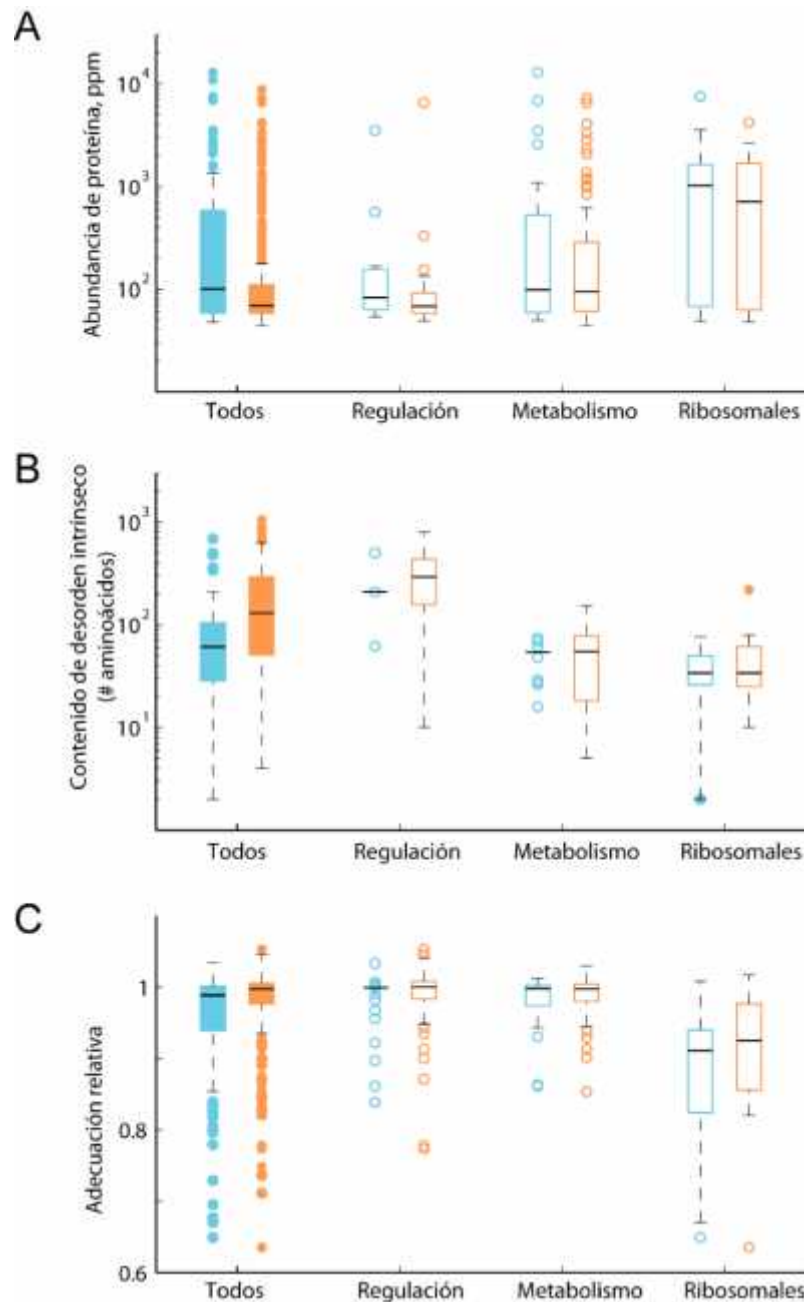


Figura 1.7 El desorden intrínseco, la abundancia de proteína y la contribución a la adecuación afectan la tasa de evolución después de la WGD pero son características de la clase funcional. Mostramos diagramas de caja para datos de *S. cerevisiae* de (A) abundancia de proteínas en ppm (Breker et al., 2013); (B) contenido aminoácidos que producen desorden intrínseco (cantidad de aminoácidos (Landry et al., 2009)); (C) adecuación relativa de mutantes sencillas (Breslow et al., 2008) de todos los genes WGD y para diferentes categorías funcionales. En azul se indican ortogrupos que no cambian su tasa de evolución después de la WGD mientras que en naranja se presentan ortogrupos que cambian su tasa significativamente.

Para explorar otras características funcionales asociadas con cambios en las tasas de evolución después de la WGD probamos el contenido de aminoácidos intrínsecamente desordenados ya que se ha observado que las proteínas desordenadas evolucionan más rápido que las proteínas ordenadas (Brown et al., 2002). En efecto, en *S. cerevisiae* las proteínas que aceleran su tasa después de la duplicación tienen en promedio un mayor contenido de aminoácidos desordenados comparadas con el grupo de proteínas con tasas constantes (**Figura 1.7-B**; $p < 10^{-4}$, test de rangos de Wilcoxon). Nuevamente cuando observamos los diferentes grupos funcionales individualmente no observamos diferencias significativas entre los grupos que cambian y los grupos que no cambian su tasa de evolución. Cabe señalar que sí se observaron diferencias significativas entre grupos funcionales. Las proteínas regulatorias tienen en promedio un mayor contenido de aminoácidos desordenados ($p < 10^{-5}$). Mientras que las enzimas metabólicas como las proteínas ribosomales tienen un menor contenido ($p < 10^{-5}$ y $p < 10^{-6}$, respectivamente). Esto sugiere fuertemente que el contenido de aminoácidos desordenados es una propiedad de la categoría funcional, como es de esperarse, por lo que no está directamente asociada con el cambio en las tasas de evolución después de la WGD.

Por otra parte, se ha observado que los genes esenciales e indispensables usualmente evolucionan a tasas menores que los genes dispensables (Hirsh & Fraser, 2001; Pal et al., 2001). Para realizar esta comparación se tomaron datos genómicos disponibles en alta resolución para mutantes individuales de pérdida de función en *S. cerevisiae*. En concordancia con observaciones previas, encontramos que los ortogrupos con tasa de evolución constante contienen parálogos con una contribución más alta a la adecuación (una adecuación menor cuando se pierden) que los parálogos con tasa de evolución cambiante (**Figura 1.7-C**; $p < 10^{-$

5, test de rangos de Wilcoxon). Nuevamente, esta tendencia parece depender más de la función; ya que no se observan diferencias significativas cuando miramos dentro de cada categoría funcional, con la excepción de las proteínas reguladoras que son marginalmente diferentes ($p=0.04$). Como es de esperarse, las proteínas ribosomales de *S. cerevisiae* contribuyen fuertemente a la adecuación y mantienen tasas de evolución constantes en las ramas pre y post-WGD. Esto sugiere que la relación entre la importancia de la función génica y el grado de compensación funcional entre parálogos determinan la relación entre la contribución a la adecuación y la tasa de evolución después de la duplicación.

En conjunto, éstos resultados sugieren que el cambio en la tasa de evolución después de la WGD está asociada con ciertas características de la proteína como son abundancia, cantidad de regiones desordenadas o contribución a la adecuación. Sin embargo, todas estas características están asociadas fuertemente con la función.

1.4.4 Una buena parte de los ortogrupos experimenta una aceleración fuerte inmediatamente después de la duplicación seguida por un periodo de estabilización

Se ha sugerido que después de la duplicación los parálogos experimentan un breve periodo de aceleración en sus tasas de evolución seguido por un periodo prolongado de estabilización (Lynch & Conery, 2000; Pegueroles, Laurie, & Albà, 2013; Scannell & Wolfe, 2008). Con esto en mente y aprovechando la habilidad de calcular tasas para ramas específicas de la filogenia se probó un segundo modelo con tres tasas de evolución diferentes: en las ramas pre-WGD (ω_0), las ramas que siguen inmediatamente post-WGD (ω_1) y las ramas modernas post-WGD (ω_2), ver **Figura 1.8-A**. Una fracción menor de ortogrupos (41.3%) fue consistente con este escenario, en el que las tasas de evolución inmediatamente después de la duplicación son diferentes a las ramas modernas. La tasa de evolución en las ramas que

siguen inmediatamente después de la duplicación es mucho más alta que la tasa de evolución pre-WGD, ver **Figura 1.8-B**. Más aún, la media de las tasas de evolución de estas ramas inmediatas está centrada alrededor de 1.3, lo que sugiere fuertemente que muchos de estos ortogrupos estuvieron sometidos a selección positiva en el periodo que siguió inmediatamente después de la WGD. De manera interesante, las ramas parálogas modernas reducen nuevamente su tasa de evolución, pero rara vez regresan al nivel pre-WGD.

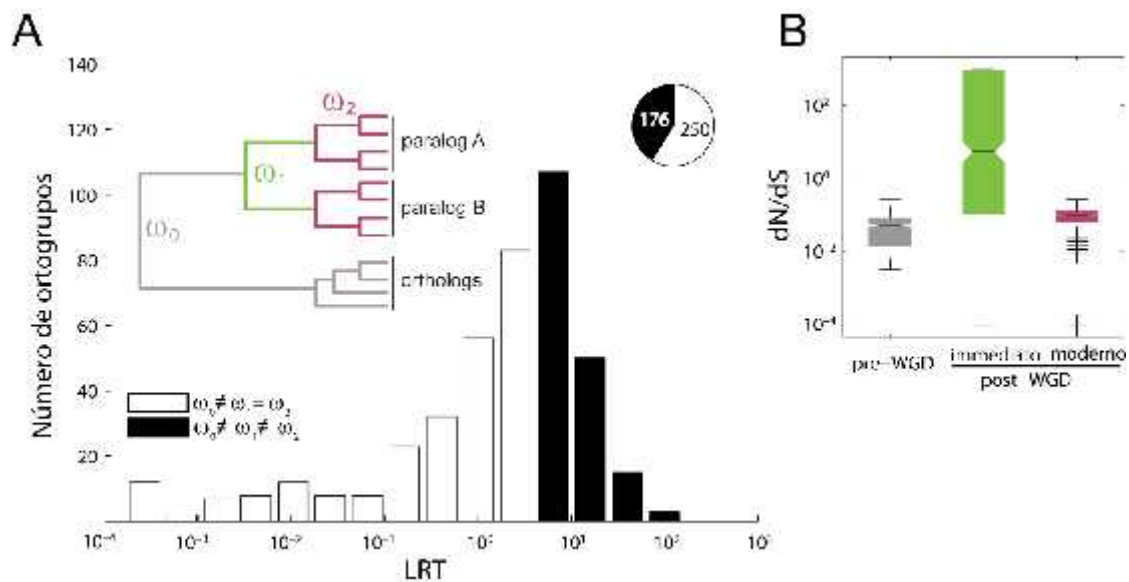


Figura 1.8 Hay una aceleración inmediata en las tasas de evolución después de la WGD seguida por una estabilización (A) Distribución de frecuencias de los valores de los test razón de verosimilitud (LRT) para la segunda prueba de hipótesis. Las barras negras indican ortogrupos para los cuales la hipótesis nula, de una sola tasa dN/dS para todas las ramas post-WGD es rechazada ($\omega_1 \neq \omega_2$); las barras vacías indican ortogrupos para los cuales la hipótesis nula no pudo ser rechazada. (B) Graficas caja para todas las tasas dN/dS de los linajes pre-WGD (ω_0), inmediatamente post-WGD (ω_1) y moderno post-WGD (ω_2).

1.4.5 Algunos ortogrupos muestran cambios asimétricos en sus tasas de evolución después de la duplicación

Se ha reportado en varios trabajos que los parálogos frecuentemente evolucionan de manera asimétrica. Nos preguntamos si un modelo con tasas de evolución asimétricas entre clados parálogos se ajusta mejor a la historia evolutiva de los ortogrupos-WGD estudiados. Con este

propósito, se diseñó una nueva prueba de hipótesis para preguntar si las tasas de evolución son diferentes entre clados parálogos. Específicamente, probamos si hay diferentes tasas de evolución entre las ramas pre-WGD (τ_0), inmediatamente después de la duplicación (τ_1) y para cada uno de los clados parálogos (τ_2 y τ_3), ver **Figura 1.9-A**. Encontramos que 52.6% ortogrupos concuerdan con este escenario de tasas asimétricas de evolución después de la aceleración inicial inmediatamente después de la WGD (**Figura 1.9-A**). Esto sugiere que muchos de los genes que aceleran sus tasas de evolución inmediatamente después de la duplicación divergen después de manera asimétrica. Adicionalmente, observamos que las proteínas regulatorias tienen una fracción más alta de parálogos con tasas de evolución asimétricas, aunque esta diferencia no es significativa (**Figura 1.9-B**). Como es de esperarse las proteínas ribosomales están sub-representadas en este escenario de tasas asimétricas debido a las restricciones estructurales y funcionales a las que están sometidas.

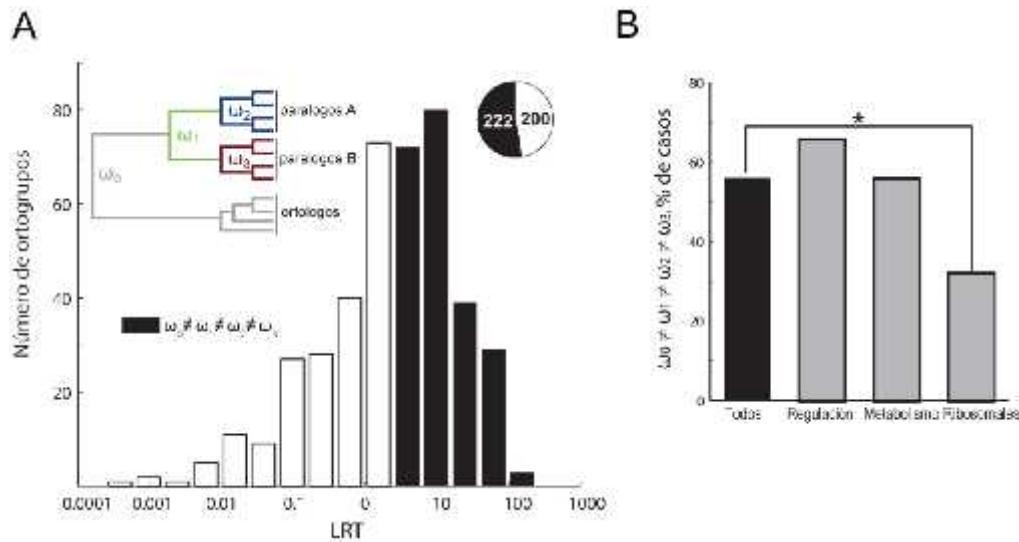


Figura 1.9 La mitad de los ortogrupos presentan tasas de evolución asimétricas después de la WGD (A) Distribución de frecuencias de los valores de los test razón de verosimilitud (LRT) para la tercera prueba de hipótesis. Las barras negras indican ortogrupos para los cuales la hipótesis nula, de una sola tasa dN/dS para los dos clados parálogos post-WGD es rechazada ($\omega_0 = 1, \omega_1 = 2, \omega_2 = 3$); las barras vacías indican ortogrupos para los cuales la hipótesis nula no pudo ser rechazada. (B) Fracción de ortogrupos para los que la hipótesis alternativa ($\omega_0 = 1, \omega_1 = 2, \omega_2 = 3$) es aceptada; la gráfica muestra todos los ortogrupos ensayados y tres categorías funcionales diferentes. Se muestran las categorías funcionales enriquecidas en el set de ortogrupos para los que se acepta la hipótesis alternativa, $*p < 0.05$ calculados con el test χ^2 .

Luego, se evaluó si la abundancia de proteína puede explicar la diferencia entre las tasas de evolución de los parálogos. Para esto, nuevamente usamos datos de *S. cerevisiae* para comparar las abundancias de proteínas parálogo *rápido* con la del parálogo *lento* (de los 222 ortogrupos con tasas de evolución asimétricas). Los resultados indican que hay una ligera tendencia a que el parálogo *lento* sea más abundante, pero en general la distribución de las abundancias relativas de las parejas de parálogos está centrada alrededor de uno (lo que significa que ambos parálogos están expresados a niveles similares); e incluso, en algunos casos el parálogo *rápido* es incluso más abundante (**Figura 1.10-A**). Sin embargo, esta tendencia no se conserva para las diferentes categorías funcionales, observamos que las enzimas metabólicas no son consistentes con la tendencia esperada; solo en un par de casos las proteínas más abundantes presentan tasas de evolución más *lentas* (**Figura 1.10-B**). Las

proteínas ribosomales se expresan a altos niveles independientemente si son parálogos rápidos o lentos; mientras que las proteínas regulatorias se expresan a bajos niveles. Por lo tanto, las diferencias en expresión no son suficientes para explicar las tasas asimétricas entre parálogos para la mayoría de los ortogrupos estudiados.

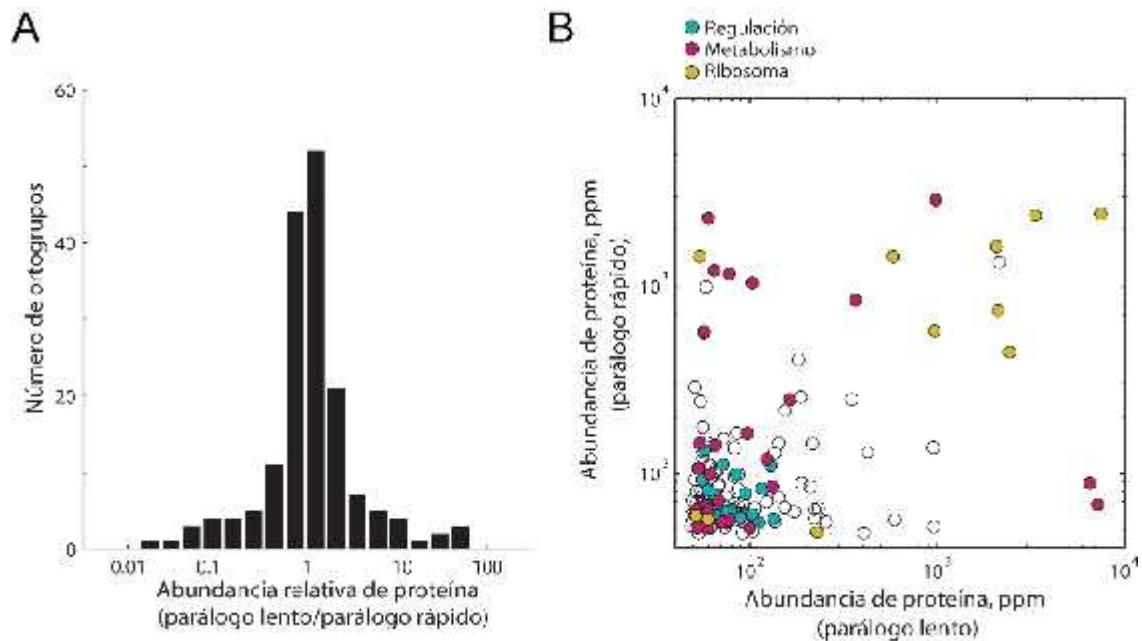


Figura 1.10 La asimetría en las tasas de evolución de los parálogos WGD no está relacionada con expresión diferencial (A) El histograma muestra la abundancia de proteína del parálogo lento relativa a la abundancia del parálogo rápido, para todos los genes WGD de *S. cerevisiae*. (B) La gráfica de dispersión muestra la abundancia de proteína absoluta del parálogo rápido comparada con la del parálogo lento. La categoría funcional de cada pareja de parálogos está señalada en círculos color azul, magenta y amarillo para las anotaciones regulatoria, metabólica y ribosomal respectivamente.

1.5 DISCUSIÓN

En este capítulo aprovechamos herramientas bioinformáticas y una base de datos que contiene ortogrupos de ohnólogos en ascomicetos, buscando caracterizar la dinámica evolutiva antes y después de la duplicación de parálogos WGD. Primero, se evaluó si las tasas de evolución antes y después de la duplicación son diferentes, encontramos que para la gran mayor parte de los ortogrupos estudiados (70%) ambas tasas son significativamente diferentes. Para determinar la dirección de éste cambio contrastamos directamente la tasa pre- con la tasa pos-WGD e inmediatamente observamos una tendencia general hacia un aumento, lo cual indica que la mayoría de estos parálogos aceleran su tasa evolutiva después de la duplicación. Ambas observaciones coinciden con múltiples trabajos previamente publicados (Davis & Petrov, 2004; Kondrashov et al., 2002; Lynch & Conery, 2000; Scannell & Wolfe, 2008).

De manera interesante, hay un grupo de genes sometidos a una selección purificadora intensa antes de la WGD que tienden a acelerar de manera drástica su tasa de evolución después de la duplicación. Debido éste grupo de genes esta enriquecido en proteínas regulatorias, como cinasas y factores de transcripción, decidimos explorar la relación de diferentes categorías funcionales con los cambios observados en las tasas de evolución despues de la WGD. Escogimos las tres categorías funcionales mejor representadas en los genes WGD: regulatorios (cinasas y factores de transcripción), proteínas ribosomales y metabolismo. Mientras que los genes metabólicos se comportan de manera similar al resto de los genes. Los genes con funciones regulatorias son mas propensos a cambiar su tasa de evolución después de la duplicación, solo una proporción menor de las proteínas ribosomales experimentan dicho cambio. La aceleración en tasa evolutiva de genes regulatorios despues

de una WGD ha sido observada previamente usando sets de datos diferentes (Corrochano et al., 2016; Kellis et al., 2004). Estos trabajos, en conjunto con nuestras observaciones sugieren que la WGD juega un papel importante en la amplificación y diversificación de las redes de regulación en eucariotes.

Los parálogos que aceleran su tasa evolutiva después de la WGD comparten ciertas propiedades moleculares en común. Las proteínas que codifican son menos abundantes, contribuyen menos a la adecuación y presentan mayor contenido de desorden intrínseco comparados con genes en ortogrupos que con tasas de evolución constantes. Sin embargo, cuando dividimos al set de ortogrupos en las categorías funcionales anteriormente usadas (regulación, metabolismo y ribosomales), no hay diferencias significativas dentro de las categorías en ninguno de los tres casos. Como es de esperarse, hay diferencias significativas entre categorías funcionales debido a que la abundancia, contribución a la adecuación y el desorden intrínseco son propiedades altamente relacionadas con la función de las proteínas. Por ejemplo, debido a la función que desempeñan, en promedio las proteínas ribosomales son más abundantes, tienen menos regiones desordenadas y contribuyen más a la adecuación que proteínas en las otras categorías funcionales. Estas observaciones en conjunto sugieren fuertemente que la función de la proteína es un determinante y primordial en la dinámica evolutiva después de una duplicación de genoma.

Luego se estudió detalladamente la dinámica evolutiva después de la duplicación, con particular interés en las ramas que divergen inmediatamente después de la duplicación. Para esto, se evaluó qué tan frecuentemente hay un cambio en la tasa de evolución en las ramas posteriores a la WGD. Los resultados indican que solo una fracción (41.3%) de los ortogrupos experimentan un cambio en la tasa de evolución después de la aceleración inicial.

Se observó una diferencia drástica en las tasas evolutivas en las ramas inmediatas post-WGD y los clados modernos, lo cual indica que para este grupo de genes hay una ráfaga de aceleración inmediata después de la duplicación seguida por una desaceleración estable en las secuencias modernas. Esto sugiere que muchos de los genes parálogos experimentan una aceleración momentánea en sus tasas de evolución, inmediatamente después de la duplicación, pero este fenómeno permanece sólo por periodos cortos de tiempo.

Finalmente, se evaluó cuantos parálogos evolucionan a tasas de diferentes entre sí después de la duplicación. Para lo cual planteamos una prueba de asimetría, y descubrimos que más de la mitad de los ortogrupos estudiados (52.6%) presentan tasas significativamente asimétricas después de la WGD. Dicha asimetría no está correlacionada con asimetría en otras propiedades moleculares como la abundancia de proteína; tampoco es una característica distintiva de los genes regulatorios o de metabolismo. Sin embargo, las proteínas ribosomales, que se caracterizan por estar altamente conservadas, rara vez evolucionan de forma asimétrica.

Estos resultados en conjunto muestran una aceleración en la tasa de cambio en las secuencias codificantes muy extendida en los parálogos después de la WGD. Sin embargo, con esta aproximación no es posible cuantificar el efecto fenotípico de tales cambios y tampoco evaluar la relevancia de las regiones no codificantes (i. e. *cis*-regulatorias). Con estas limitaciones en mente en la siguiente parte decidimos investigar más a fondo la relevancia fenotípica de las regiones codificantes y *cis*-regulatorias de estos genes, mediante una estrategia experimental.

CAPÍTULO 2

DIVERGENCIA FUNCIONAL DE LAS REGIONES *CIS*-REGULATORIA Y CODIFICANTE DE GENES DUPLICADOS EN LEVADURA²

Los cambios a nivel genético que modifican fenotipos son el motor de la innovación funcional y por tanto de la diversificación fenotípica en la naturaleza. La genética clásica se ha dado a la tarea de estudiar principalmente cambios en la región codificante, ya sea generando mutantes de pérdida de función o estimando los cambios en la secuencia de las proteínas. Por otro lado, a pesar de que podemos fácilmente detectar cambios en la expresión génica poco sabemos de cómo estos cambios afectan el fenotipo. Más aún, hay muy pocos trabajos que hayan desarrollado estrategias para evaluar la importancia de cambios en ambas regiones con una métrica directamente equivalente, por lo que se desconoce ampliamente la contribución de cada región a la innovación funcional. Con esta interrogante en mente, en este capítulo nos dimos a la tarea de diseñar una estrategia experimental que nos permite contrastar el efecto fenotípico de cambios en regulación y en la región codificante de un grupo pequeño de genes parálogos.

² Este capítulo contiene texto y figuras tomados o modificados del trabajo " Coupled Divergence of Coding and *Cis*-Regulatory Functions after Gene Duplication" en el Anexo G (Ascencio et al., en preparación).

2.1 ANTECEDENTES

Los genes duplicados son el modelo favorito para estudiar la evolución de nuevas funciones. La redundancia funcional que surge debido al traslape funcional entre las copias da lugar a una relajación en la presión de selección. Los factores que afectan dicho proceso fueron estudiados en nuestra sección anterior. Dicha relajación en la restricción funcional da lugar a un aumento en la cantidad de cambios puntuales sobre las secuencias que pueden afectar tanto la región codificante como la *cis*-regulatoria. La acumulación de mutaciones en la región codificante de los parálogos puede modificar sitios activos, función bioquímica o especificidad en las interacciones con otras proteínas (Gagnon-Arsenault et al., 2013). Mutaciones en las regiones no codificantes pueden causar pérdida o reconfiguración de las interacciones regulatorias provocando divergencia en expresión entre los parálogos (Guan et al., 2007; Hittinger & Carroll, 2007; Teichmann & Babu, 2004).

La divergencia en expresión, inferida con datos de microarreglos en *S. cerevisiae* y *Drosophila melanogaster*, es más extensa entre genes duplicados comparada con el del resto del genoma (Gu, Nicolae, Lu, & Li, 2002). Se ha sugerido que la cantidad de elementos regulatorios compartidos entre duplicados se reduce con el tiempo debido a la existencia de selección positiva (B. Papp, C. Pal, & L. D. Hurst, 2003). Se ha mostrado que los parálogos con una alta similitud en secuencia (con ID>90%) exhiben una divergencia regulatoria más evidente que parálogos con mayor divergencia en secuencia (Tirosh & Barkai, 2007).

Se estima que el 80% de los cambios adaptativos en el genoma son elementos regulatorios, mientras que solo el 20% son en región codificante de genes (Halligan et al., 2013). Pero esto

no necesariamente implica que los cambios en regiones no codificantes dirigen la adaptación, porque el efecto adecuacional de las sustituciones en la región codificante pueden ser mayor que las sustituciones en regiones no codificantes. De hecho, se ha calculado que el efecto promedio en la adecuación de una mutación en un elemento regulatorio es mucho menor comparada con una mutación no sinónima en la región codificante (Halligan et al., 2013). Esto sugiere que la mayoría de los cambios que modifican la adecuación son en la región codificante de los genes.

Mediante una aproximación experimental, se abordará la siguiente pregunta: ¿cuál es el papel relativo de las regiones cis-regulatorias y las regiones codificantes en la divergencia fenotípica de los genes duplicados? Para dar respuesta llevaremos a cabo un análisis de complementación genética en el que de manera sistemática evaluamos el índice de divergencia de cada región cuantificando directamente su efecto sobre la adecuación.

2.2 HIPÓTESIS

La divergencia funcional de genes duplicados es dirigida tanto en cambios de proteína como en regulación. Sin embargo, se desconoce la contribución de cada región a la adaptación. Existe amplia evidencia de que la regulación diverge rápidamente después de una duplicación. No obstante, el efecto fenotípico de dicha divergencia es ampliamente desconocido. Aun cuando la secuencia codificante evoluciona más lentamente se ha sugerido que éstos pocos cambios afectan más frecuentemente la adecuación que los cambios en la región *cis*-regulatoria.

2.3 MÉTODOS

2.3.1 Cepas y plásmidos usados.

Tomamos mutantes individuales sencillas de 711 genes duplicados en la WGD de la colección de *deleciones* de *S. cerevesiae* (Giaever et al., 2002). Cruzamos dicha colección con la cepa Y7092-*YFP-natR* (Tong & Boone, 2006) y seleccionamos productos meióticos haploides para obtener un set de mutantes con el siguiente genotipo *MATa ho ::Venus-natR can1 ::STE2pr-SP_his5 lyp1 his3 1 leu2 0 ura3 0 met15 0 LYS2+*, ver **Tabla 2.1**.

Se introdujo el módulo CFP-*caURA3-MX4*, que confiere prototrofia a uracilo, en la cepa parental *BY4741* mediante remplazo por recombinación dirigida por PCR (Baudin, Ozier-Kalogeropoulos, Denouel, Lacroute, & Cullin, 1993) en el locus neutral *HO*. La cepa diploide inicial *YDA100* se obtuvo mediante la cruce de las cepas *BY4741-CFP::URA3* y *Y7092-YFP-natR*. El plásmido de clonación fue el *pFA6-KANMX4* que incluye un marcador de selección para resistencia a geneticina (G418) (Wach, Brachat, Pohlmann, & Philippsen, 1994).

El medio de selección de diploides está compuesto por base nitrogenada YNB (*BD Difco*) sin uracilo, 2% glucosa, glutamato monosódico (*Sigma Aldrich*) como fuente de nitrógeno con 100mg/L de clonNAT (*Werner BioAgents*). El medio de competencia es una modificación de YNB de baja fluorescencia (Sheff & Thorn, 2004) con 2% de glucosa y suplementado con una mezcla completa de aminoácidos (CSM, *MP Biomedicals*). En el caso del medio adicionado con doxiciclina, el mismo medio de competencia se suplementó con quince concentraciones desde 2.5 mg/L diluciones 1:2 hasta 0.0002 mg/L.

Tabla 2.1. Cepas de *Saccharomyces cerevisiae* usadas en este estudio

Cepa	Genotipo	Origen	Notas
Y7092	<i>MAT can1 ::STE2pr-SP_his5 lyp1 his3 1 leu2 0 ura3 0 met15 0 LYS2+</i>	(Tong & Boone, 2006)	Cepa parental SGA
R1158	<i>MATa URA3::CMV-tTA his3 1 leu2 0 met15 01</i>	(T. R. Hughes et al., 2000)	Sistema <i>tet-off</i>
Y7092-YFP	<i>MAT ho ::Venus-natR can1 ::STE2pr-SP_his5 lyp1 his3 1 leu2 0 ura3 0 met15 0 LYS2+</i>	(Wach et al., 1994)	Desciende de Y7092
Y7092-CFP	<i>MAT ho ::Cerulean-natR can1 ::STE2pr-SP_his5 lyp1 his3 1 leu2 0 ura3 0 met15 0 LYS2+</i>	(Wach et al., 1994)	Desciende de Y7092
Y7092-YFP xxx	<i>MAT ho ::Venus-natR can1 ::STE2pr-SP_his5 lyp1 his3 1 leu2 0 ura3 0 met15 0 LYS2+ ::kanR</i>	(Wach et al., 1994)	Cepas xxx
BY4741- CFP	<i>MATa ho ::Cerulean-URA3 his3 1 leu2 0 lys2 0 ura3 0</i>	éste trabajo	Desciende de BY4741
YDA100	<i>MATa/ ho ::Cerulean-URA3/ho ::Venus-natR his3 1/his3 1 leu2 0/leu2 0 met15 0/met15 0 ura3 0/ura3 0</i>	éste trabajo	Cruza de BY4741-CFP y Y7092-YFP
R1158-CFP	<i>MATa ho ::Venus-natR URA3::CMV-tTA his3 1 leu2 0 met15 01</i>	éste trabajo	Desciende de R1158

2.3.3 Construcción de cepas quiméricas y controles.

Cada uno de los genes seleccionados se amplificó usando DNA genómico de la cepa *S288C* como templado y fue clonado en el vector *pFA6-KanMX4*, con los oligonucleótidos descritos en el **Anexo C**. Se usó un solo sitio de restricción (*XmaI*) para la clonación y se seleccionaron construcciones con inserciones en ambas direcciones. La clona en dirección 5'-3' se usó para remplazo del ORF y se amplificó con oligonucleótidos híbridos con F1 y R3 fusionados a 40bp de homología al inicio y final del marco de lectura de cada gen respectivamente. Éstas secuencias amplificadas se usaron para remplazar las regiones correspondientes del gen parálogo mediante recombinación dirigida por PCR en el locus genómico. Los plásmidos con inserciones 3'-5' se usaron para remplazar el promotor con un primer sentido F1 con 40 bases de homología en promedio a -600pb al inicio de la transcripción y en contrasentido con R3 fusionado a -1 a +45pb del inicio de la transcripción de cada gen (ver **Anexo D**, oligonucleótidos de integración usados). Dicha estrategia es ilustrada en la **Figura 2.1**.

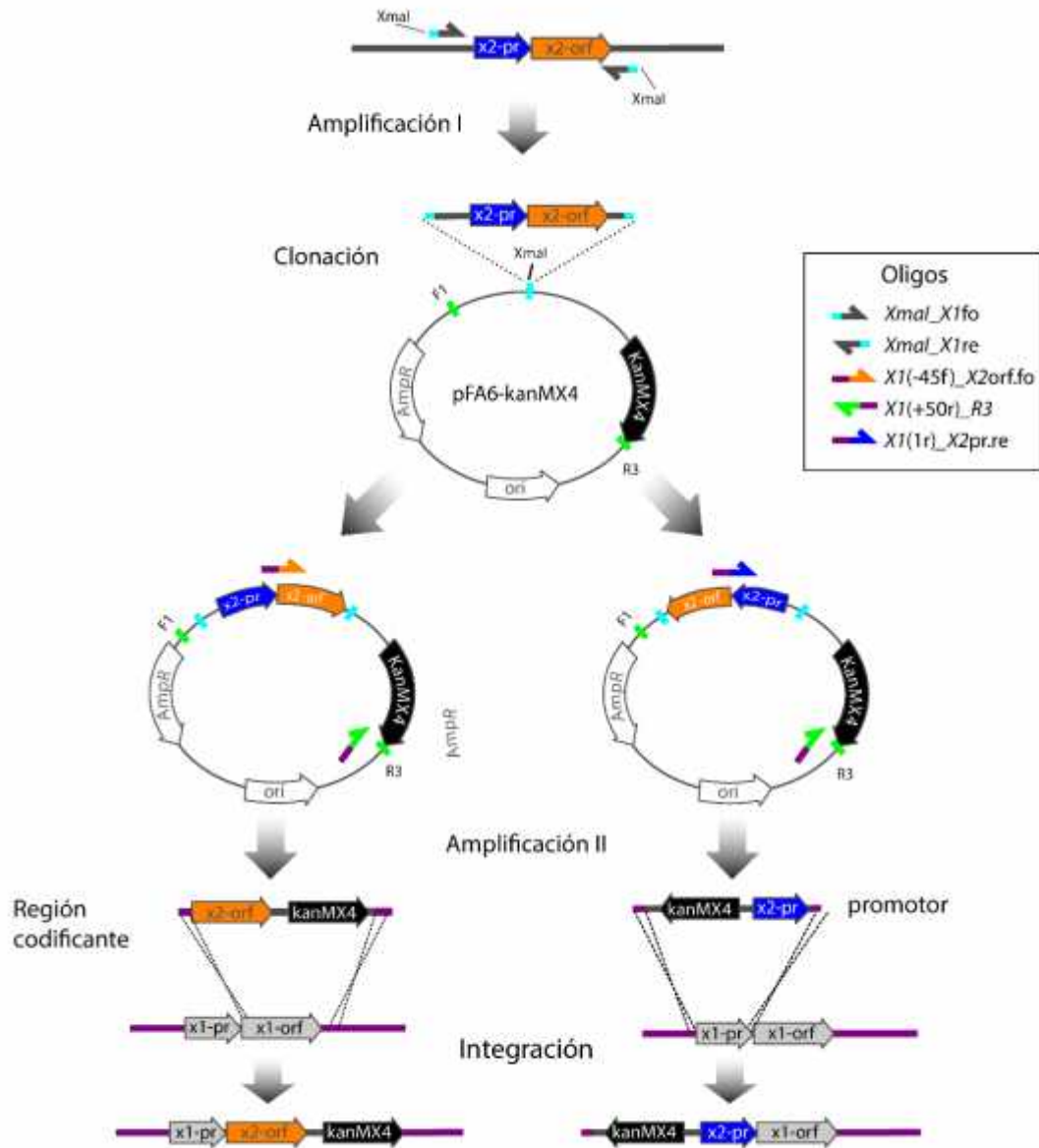


Figura 2.1 Estrategia experimental para intercambiar regiones entre parálogos. El esquema representa los pasos que seguimos para hacer las construcciones quiméricas. Ambos genes con sus respectivas regiones genéticas (promotor y ORF) se clonan en plásmidos con marcadores de selección. Cada región es amplificada junto con el marcador de selección (*KanMX4*) para posteriormente ser integrada en el locus destino, mediante recombinación homóloga. El resultado es una construcción quimérica, el promotor nativo expresando a la proteína paróloga (izquierda), o el promotor parólogo expresando la proteína paróloga (derecha).

2.3.4 Calibración de condiciones usadas en la competencia

Para seleccionar las condiciones para suplementar el medio se realizaron cinéticas de crecimiento con la cepa parental Y7092 en diferentes concentraciones de etanol, NaCl, peróxido, cafeína y ácido acético en un gradiente de concentraciones mostradas en la **Figura 2.2**. En cada cinética se inoculó con un cultivo saturado de 48hr de la cepa parental en una proporción 1:16. Durante la cinética se realizaron lecturas cada hora durante 24 horas. Para cada condición seleccionamos la concentración (marcada en negritas en la **Figura 2.2**) que provoca una disminución importante pero no total del crecimiento.

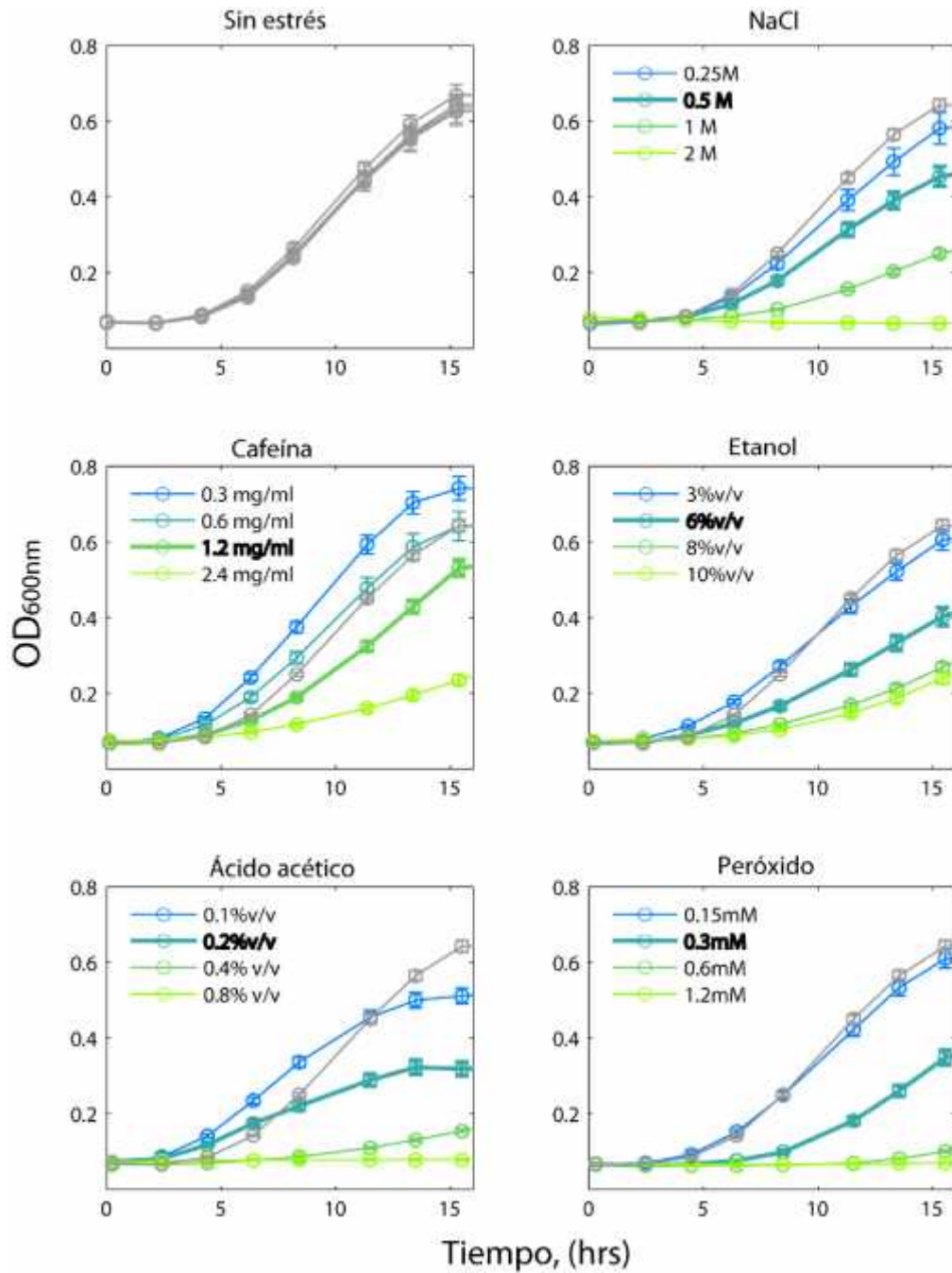


Figura 2.2 Calibración de las diferentes condiciones usadas en la competencia. Curvas de crecimiento de la cepa de referencia creciendo en diferentes condiciones. En cada caso adicionamos las distintas concentraciones del agente estresante que se muestran en diferentes condiciones. En negritas la concentración elegida para los ensayos posteriores. En negritas mostramos la concentración seleccionada para los siguientes experimentos.

2.3.5 Construcción de cepas de promotor regulable *tetO₇*

Para la construcción de cepas con promotor regulable se usó la cepa parental *R1158-CFP-URA3* (ver **Tabla 2.1**) que expresa el elemento transactivador (*tTA*) necesario para el funcionamiento del sistema de expresión regulable *tetO₇* (Yen, Gitsham, Wishart, Oliver, & Zhang, 2003). Sustituimos el promotor nativo de cada gen con el promotor regulable *tetO₇* mediante recombinación dirigida por PCR en el locus original. Para este propósito, se emplearon oligonucleótidos híbridos (**Anexo D**) con homología al plásmido *pAG32* (Yen et al., 2003) y a al promotor del gen objetivo (-600pb y -1 a -45pb del inicio de la transcripción). Se utilizó el módulo *hph*, que confiere resistencia a *higromicina B*, como marcador de selección. Las mutantes positivas se seleccionaron en YPAD con 300mg/L de higromicina B (Invitrogen) y 100mg/L nourseotricina (Werner BioAgents).

2.3.6 Método de competencia automatizado.

El método de competencia usado es una modificación del método usado por De Luna *et al.* (2008). Partimos de cultivos saturados de 48hrs y se realizó una mezcla de cultivos con 50µl de la mutante YFP con 100ul de la cepa referencia Y7092-CFP, en microplacas de 96 pozos (*Costar*). Las microplacas con 150µl de medio de competencia se inocularon con 10µl de los cultivos saturados previamente mezclados. El cultivo se diluyó cada 24 horas en medio fresco de competencia en una proporción 1:16 mediante el sistema automatizado *Freedom EVO* (*Tecan Ltd*) acoplado a un multilector (Infinite Reader M100, *Tecan Ltd*) para microplacas. Se realizaron lecturas cada 2hrs de OD600nm, fluorescencia *YFP* (515/528nm con ganancia ajustada a 130) y *CFP* (433/475nm con ganancia ajustada a 145). El equipo y los cultivos fueron mantenidos a 30°C y ~70% de humedad relativa. El experimento se siguió durante 6 días.

2.4 RESULTADOS

2.4.1 Un 10% de las parejas de duplicados WGD presentan una divergencia fenotípica asimétrica.

Para encontrar parejas de parálogos con divergencia observable a nivel fenotípico cuantificamos la adecuación relativa de una colección de cepas de 711 mutantes sencillas de pérdida de función de genes duplicados en la WGD en *S. cerevisiae*. Para dicha cuantificación usamos un método de competencia de alta resolución que resultó tener una alta replicabilidad (ver **Figura 2.3**). Ordenamos cada pareja de genes de acuerdo a su contribución a la adecuación, nombramos al gen con mayor contribución como *X1* y al gen con una contribución menor como *X2*.

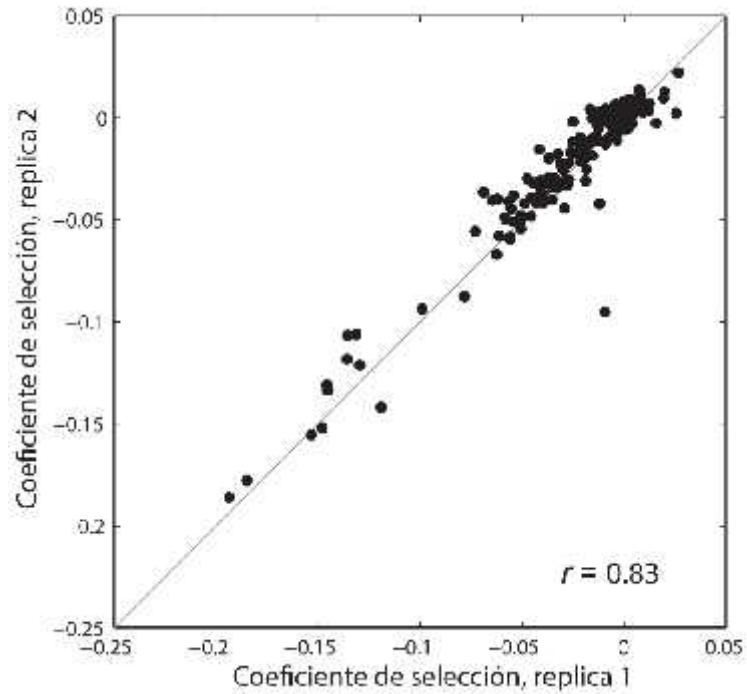


Figura 2.3 Replicabilidad técnica del ensayo de competencia para determinar adecuación relativa. En la figura comparamos el coeficiente de selección de las dos replicas técnicas incluidas en cada ensayo de competencia.

Nuestros resultados sugieren que la mayor parte de los parálogos se compensan funcionalmente por lo que ninguna de sus mutantes de pérdida de función exhibe una disminución en la adecuación (**Figura 2.4**).

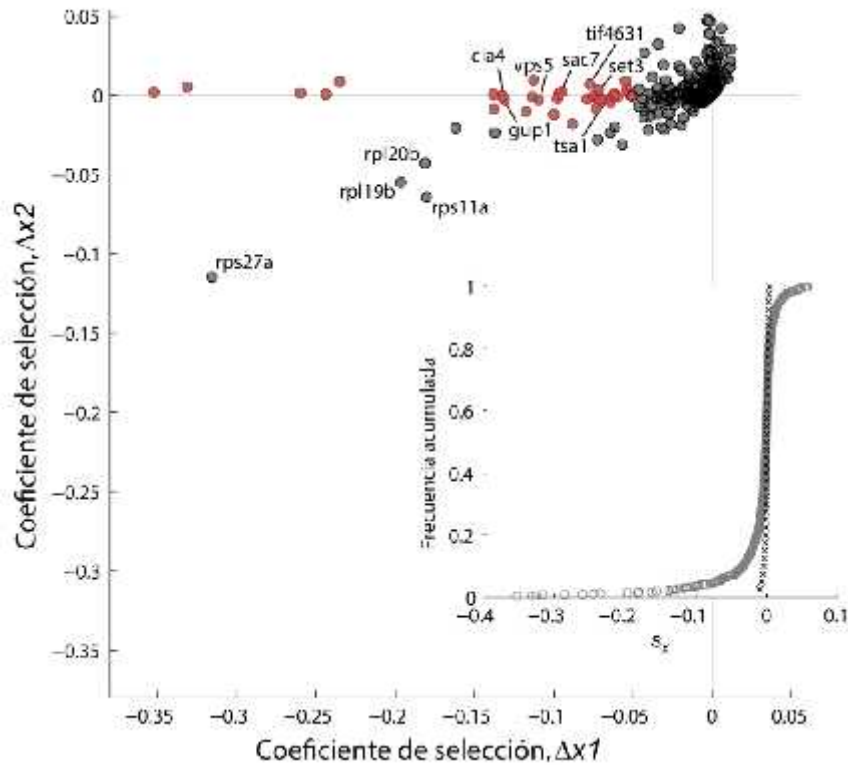


Figura 2.4. Comparación de la adecuación de las mutantes de pérdida de función de los parálogos WGD. Comparamos la adecuación relativa de la mutante sencilla de cada gen con la de su parálogo, para todas las parejas WGD. En cada pareja ordenamos a los genes por efectos, el gen como mayor impacto en la adecuación se denomina X1, mientras que el parálogo con menor efecto le denominamos X2. En la parte inferior derecha mostramos una gráfica de frecuencia acumulada de todos los coeficientes de selección cuantificados en el experimento. Los círculos representan las mutantes sencillas de pérdida de función, mientras que las “x” representan las referencias.

Las proteínas ribosomales se encuentran en un grupo especial, ya que ambos integrantes de la pareja de duplicados contribuyen de manera similar a la adecuación. Se ha sugerido que estas proteínas tienen una alta demanda durante el crecimiento por lo que son requeridas en dosis altas, por lo que la disminución de la cantidad de cualquiera de las dos proteínas resulta en una disminución en la velocidad de crecimiento (Balazs Papp, Csaba Pal, & Laurence D Hurst, 2003). Además, dichas proteínas están altamente conservadas en secuencia y regulación, por lo que se cree que son un caso clásico de selección positiva por dosis.

Identificamos un grupo de 31 (~10%) parejas de genes duplicados en los que solo un integrante es indispensable (*X1*), mientras que su pareja (*X2*) es dispensable. Esto significa que el parólogo *X2* no es capaz de compensar funcionalmente la ausencia del parólogo *X1*. Estos fenotipos asimétricos indican una pérdida de traslape funcional entre los parálogos; donde, diferencias en sus propiedades bioquímicas o *cis*-regulatorias provocan que los genes dispensables sean incapaces de compensar la ausencia de los parálogos indispensables en las condiciones ensayadas. Este grupo de genes es un modelo ideal para estudiar el papel de la región *cis*-regulatoria y codificante, por lo que seleccionamos un grupo de ellos para evaluar el papel de ambas regiones en dicha divergencia funcional (**Tabla 2.2**). Nuestro set de genes incluye parejas de duplicados con diferentes grados de conservación de secuencia.

2.4.2 Comparación directa de la divergencia fenotípica de la región *cis*-regulatoria y codificante de genes parálogos

Inmediatamente después de la duplicación ambas copias son idénticas tanto en su región *cis*-regulatoria como codificante; por lo que la mayoría de las parejas experimentan una relajación importante en la presión de selección que previene mutaciones en sus secuencias. Como resultado, los genes parálogos acumulan mutaciones en la región regulatoria, la región codificante o ambas; que pueden dar lugar a diversificación funcional (**Figure 2.5**).

Tabla 2.2 Propiedades de los parálogos WGD usados en este estudio

<i>X1</i>	<i>X2</i>	<i>X1</i> es esencial	CAI-1 ^a	CAI-2 ^a	AA% id	Ka ^b	Descripción <i>X1</i>
<i>RHR2</i>	<i>HOR2</i>	no	0.604	0.292	92%	0.044	Enzima glicerol-1-fosfato
<i>TSA1</i>	<i>TSA2</i>	no	0.71	0.27	85%	0.086	Tiorredoxin peroxidase, actúa asociada al ribosoma y en el citoplasma
<i>ITR1</i>	<i>ITR2</i>	no	0.195	0.173	80%	0.152	Transportador de mioinositol
<i>ERV14</i>	<i>ERV15</i>	no	0.311	0.105	62%	0.389	Proteína de vesícula COPII
<i>FEN1</i>	<i>ELO1</i>	no	0.239	0.134	59%	0.371	Elongasa de ácidos grasos
<i>GUP1</i>	<i>GUP2</i>	no	0.17	0.12	55%	0.352	Proteína de membrana plasmática de remodelación de anclas GPI
<i>BUL2</i>	<i>BUL1</i>	no	0.144	0.138	53%	0.414	Componente del complejo Rsp5p E3-ubiquitin ligasa
<i>CLA4</i>	<i>SKM1</i>	no	0.13	0.11	50%	0.396	PAK cinasa de transducción de señales, activada por Cdc42p
<i>TIF4631</i>	<i>TIF4632</i>	no	0.25	0.18	50%	0.401	Factor eIF4G de inicio de la traducción
<i>RHR2</i>	<i>HOR2</i>	no	0.604	0.292	92%	0.044	DL-glicerol-3-fosfatasa
<i>CLA4</i>	<i>SKM1</i>	no	0.13	0.11	50%	0.396	PAK cinasa de transducción de señales, activada por Cdc42p
<i>TIF4631</i>	<i>TIF4632</i>	no	0.25	0.18	50%	0.401	Factor eIF4G de inicio de la traducción
<i>CDC19</i>	<i>PYK2</i>	sí	0.893	0.13	70%	0.206	Piruvato cinasa
<i>GSP1</i>	<i>GSP2</i>	sí	0.62	0.26	95%	0.013	Proteína de unión a GTP involucrada en organización nuclear, procesamiento de RNA y transporte
<i>PET9</i>	<i>AAC3</i>	sí	0.54	0.2	89%	0.063	El mayor acarreador de ADP/ATP de la membrana interna mitocondrial
<i>DED1</i>	<i>DBP1</i>	sí	0.38	0.12	69%	0.225	RNA helicasa DEAD de caja (Asp-Glu-Ala-Asp) dependiente de ATP
<i>SDH3</i>	<i>YMR118C</i>	sí	0.2	0.11	40%	0.63	Subunidad del Citocromo b de la succinato deshidrogenasa
<i>REB1</i>	<i>YDR026C</i>	sí	0.2	0.16	39%	0.53	Proteína de unión a RNA polymerasa I
<i>RSC6</i>	<i>SNF12</i>	sí	0.13	0.124	27%	0.824	Subunidad del complejo RSC de remodelación de cromatina
<i>PDI1</i>	<i>EUG1</i>	sí	0.199	0.155	39%	0.53	Proteína disulfuro-isomerasa

a. CAI del trabajo de Kellis et al, 2004

b. Ka obtenido del Yeast Genome Order Browser (<http://ygob.ucd.ie/>)

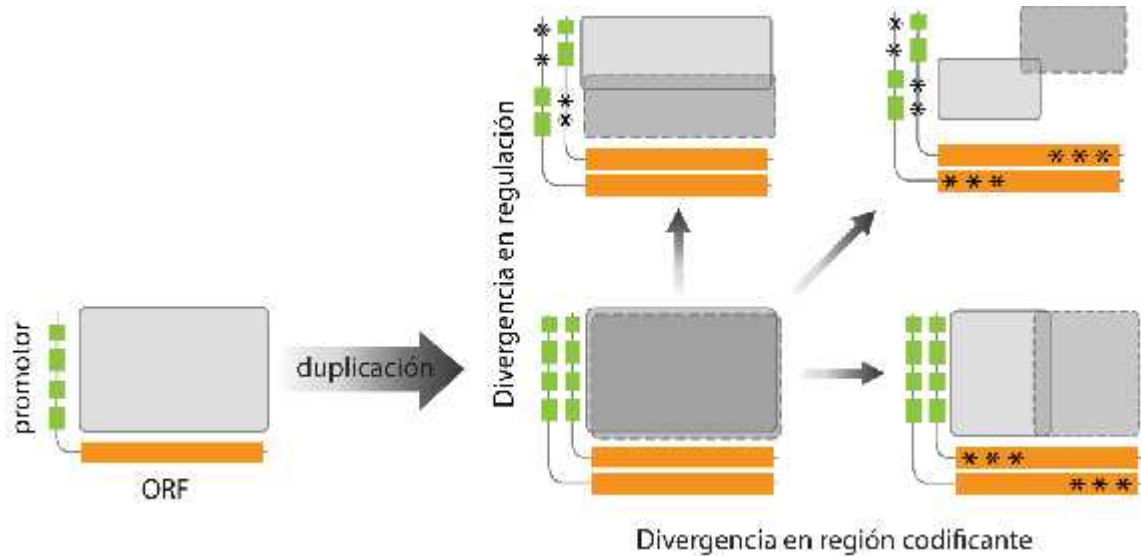


Figura 2.5 Los diferentes escenarios evolutivos de divergencia funcional de genes parálogos después de la **duplicación**. Representamos distintos elementos regulatorios en verde; en amarillo la región codificante. El cuadro gris representa el espacio funcional que depende de las dos regiones genéticas

Con la finalidad de determinar el efecto fenotípico de los cambios de cada región usamos una estrategia de complementación genética entre las regiones *cis*-regulatorias y codificantes de los parálogos, descritas en la **Figura 2.6-A**. Dicha estrategia experimental nos permitió cuantificar el grado de complementación de cada región del gen con efecto fenotípico significativo con la región correspondiente del parálogo (**Figura 2.6-B**). A continuación, comparamos la adecuación relativa de cada una de las construcciones y calculamos un *índice de divergencia fenotípica* para cada región génica; un índice de divergencia de región codificante (D_{orf}) y un índice de divergencia de promotor (D_{prm}). Éstos índices de divergencia consideran el grado de complementación de cada región, con respecto al control positivo, normalizada con la adecuación de la mutante sencilla de *X1* (**ver Figura 2.6 C**).

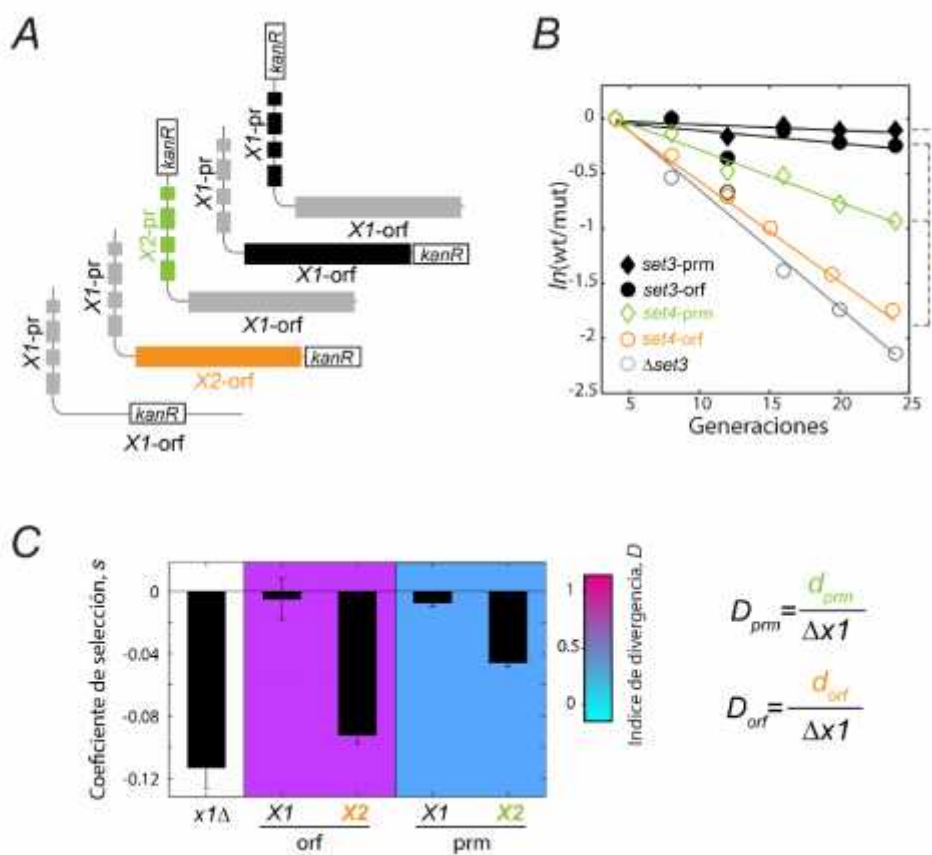


Figura 2.6 Mediante una estrategia de complementación genética cuantificamos un índice de divergencia de cada región génica. (A) Estrategia de complementación genética. Mostramos las construcciones para la complementación, todas se realizaron en el locus nativo del parálogo *X1* (el que contribuye más a la adecuación). De arriba para abajo, las primeras dos construcciones son controles positivos de remplazo de promotor y ORF; las dos construcciones en la mitad son las construcciones quiméricas de complementación paráloga de promotor y ORF; la última construcción es la mutante *X1* de pérdida de función. (B) Un ejemplo del ensayo de complementación genética de *SET3/4*. La pendiente del cociente de fluorescencias en función de las generaciones se denomina coeficiente de selección. La diferencia de los coeficientes de selección de los controles positivos y las cepas de complementación de promotor y ORF se usan para calcular los índices de divergencia D_{prm} y D_{orf} de cada región. (C) El índice de divergencia se calcula basándose en los datos de adecuación del ensayo de complementación descrito arriba. Un índice de divergencia de 1 significa que la región paróloga (*X2*) no compensa la ausencia de la región original, mientras que un índice de divergencia cercano a cero indica una falta de complementación por parte la región correspondiente de *X2*.

En la **Figura 2.7** mostramos los resultados para todas las condiciones y parejas ensayadas. Adicionalmente, los efectos no son totalmente consistentes en todas las condiciones, quizá el mejor ejemplo de ello son los parálogos *TSA1/TSA2* los cuales parecen complementarse

de manera radicalmente diferente en peróxido y muestran efectos marginales en las demás condiciones. De manera interesante, ambas regiones fallan en complementar la ausencia de la región correspondiente de *TSA1* en dicha condición.

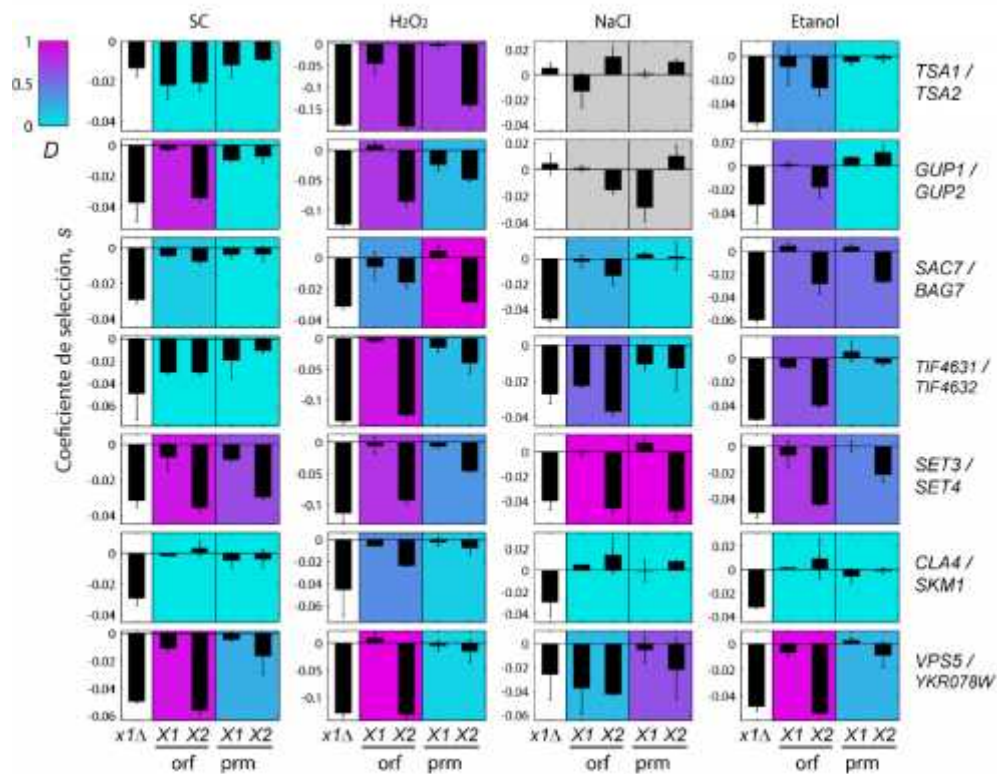


Figura 2.7 El impacto fenotípico de la divergencia codificante y *cis*-regulatoria después de la duplicación génica. Datos de adecuación para el ensayo de complementación genética para siete parejas de parálogos (filas) en cuatro condiciones diferentes (columnas). En cada gráfica presentamos cinco barras que corresponden a las cinco construcciones genéticas, las barras de error representan el error estándar de la media de tres replicas biológicas. El índice de divergencia (D_{orf} a la derecha y D_{prm} a la izquierda) es indicado con un gradiente de color, de una divergencia funcional alta (rosa) a una divergencia funcional pequeña (azul).

Es importante señalar que tuvimos que descartar un set de genes por que los controles no se ajustaban a lo que se esperaba, como es el caso de *RHR2/HOR2*, *ERV14/ERV15* y *FEN1/ELO1* por lo que es probable que existan errores durante la recombinación que

posiblemente estén alterando el marco de lectura de tales genes, ver **Figura 2.8**. En el caso de *ITR1/ITR2* no se pudieron obtener varias construcciones.

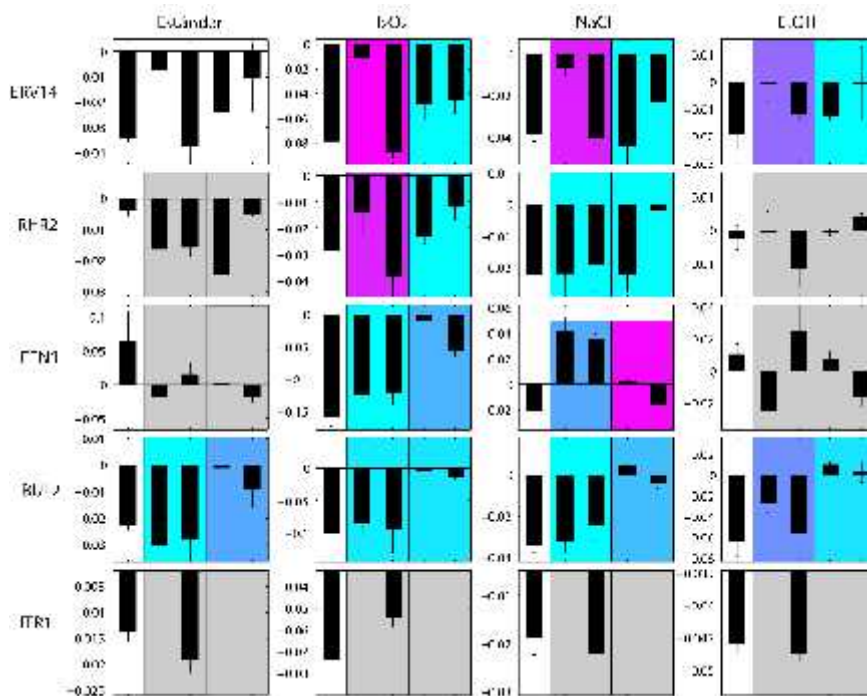


Figura 2.8 Cepas descartadas por diversas razones. Datos de adecuación para el ensayo de complementación genética para siete parejas de parálogos (filas) en cuatro condiciones diferentes (columnas). En cada gráfica presentamos cinco barras que corresponden a las cinco construcciones genéticas, las barras de error representan el error estándar de la media de tres replicas biológicas. El índice de divergencia (D_{orf} a la derecha y D_{prm} a la izquierda) es indicado con un gradiente de color, de una divergencia funcional alta (rosa) a una divergencia funcional pequeña (azul).

Resulta interesante que fue más común observar que una mayor adecuación en la construcción donde el promotor parálogo expresa la proteína endógena que cuando la proteína paróloga es controlada por el promotor endógeno. Es importante señalar que casi en todos los casos en los que observamos una fuerte divergencia del promotor esta acoplada con una divergencia fuerte en la proteína. De hecho, el grado de divergencia funcional de la región *cis*-regulatoria y codificante en las diferentes condiciones están significativamente correlacionadas (**Figura 2.9**). Esta observación se sostiene incluso cuando integramos todas

las condiciones para una pareja de genes, círculos de colores en **Figura 2.9**. Finalmente, notamos que el alto grado de divergencia funcional (D_{orf}) en la proteína no se asocia en todos los casos con una alta tasa de evolución a nivel de secuencia codificante (ver valores Ka en **Figura 2.9**).

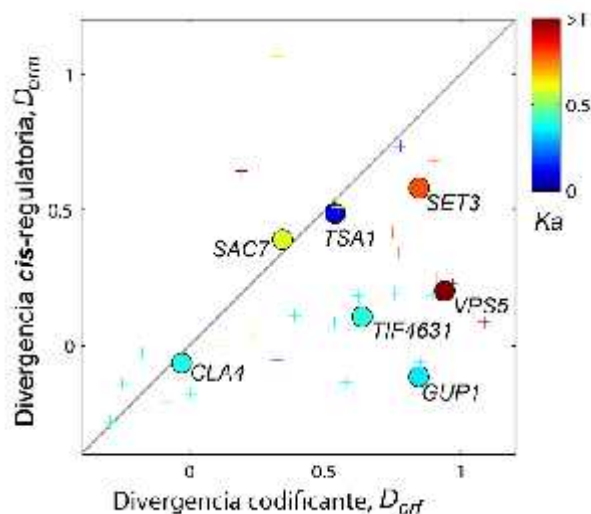


Figura 2.9 Comparación directa de la divergencia fenotípica de la región codificante y región *cis-regulatoria*. En la gráfica representamos el índice de divergencia en región codificante (D_{orf}) en función del índice de divergencia en región regulatoria (D_{prm}). Las cruces representan los índices en cada condición ensayada mientras que los círculos representan el promedio de todas las condiciones. El código de color es determinado la tasa de mutaciones no sinónimas (Ka) entre las secuencias parálogas

Nuestras estimaciones cuantitativas de divergencia D_{prm} y D_{orf} indican niveles similares de pérdida de traslape funcional de las proteínas y promotores en varias parejas de genes. Por ejemplo, encontramos que la proteína Set4 expresada por el promotor $SET3_{prm}$ no reestablece la adecuación de *set3*; tampoco el promotor $SET3_{prm}$ controlando la expresión de la proteína Set3. Se sabe que $SET3$ codifica para el miembro más importante del complejo de deacetilación de histonas Set2C (Aasland, Gibson, & Stewart, 1995), mientras que ninguna función ha sido asociada con el parálogo $SET4$, aunque conserva el dominio SET comparte poca similitud en secuencia de aminoácidos con $SET3$.

2.4.3 Las parejas de duplicados con divergencia funcional extrema exhiben divergencia total en la región codificante y cis-regulatoria

En los experimentos descritos anteriormente estudiamos un grupo de genes que ha divergido de forma que la máxima adecuación no puede ser lograda cuando el parálogo *X1* está ausente. Sin embargo, estas mutaciones no son letales, ya sea porque el parálogo *X2* es capaz de proveer compensación parcial ante la ausencia de *X1*, o porque la función *X1* no es completamente esencial para el crecimiento. Debido a que los parálogos esenciales proveen un escenario de divergencia fenotípica completa, quisimos evaluar la capacidad de compensación de las regiones *cis*-regulatoria y codificante en parálogos WGD esenciales. Con este propósito, decidimos evaluar la viabilidad de las construcciones de remplazo de promotor y región *cis*-regulatoria para siete parejas de genes con un integrante esencial. En la mayoría de los casos, ni la región codificante ni la región regulatoria son capaces de revertir el fenotipo letal ante la ausencia del parálogo esencial *X1* (**Figura 2.10**), lo que denota una fuerte divergencia funcional en ambas regiones parálogas. En conjunto estos resultados sugieren que en casos de divergencia extrema la divergencia es dirigida por la región codificante y *cis*-regulatoria de forma conjunta. Existe evidencia en la literatura que denota un alto grado de divergencia en algunas de estas parejas. Por ejemplo, *DED1* es una RNA helicasa dependiente de la caja DEAD, a pesar de que comparte una identidad de secuencia importante (69%) con su parálogo *DBP1*, ni la proteína Dbp1 ni el *DBP1_{prm}* son capaces de restaurar el crecimiento. Sorprendentemente, incluso el ortólogo de humano DDX3X (53% de identidad) sí es capaz de rescatar la mutante *ded1* (Iost, Dreyfus, & Linder, 1999).

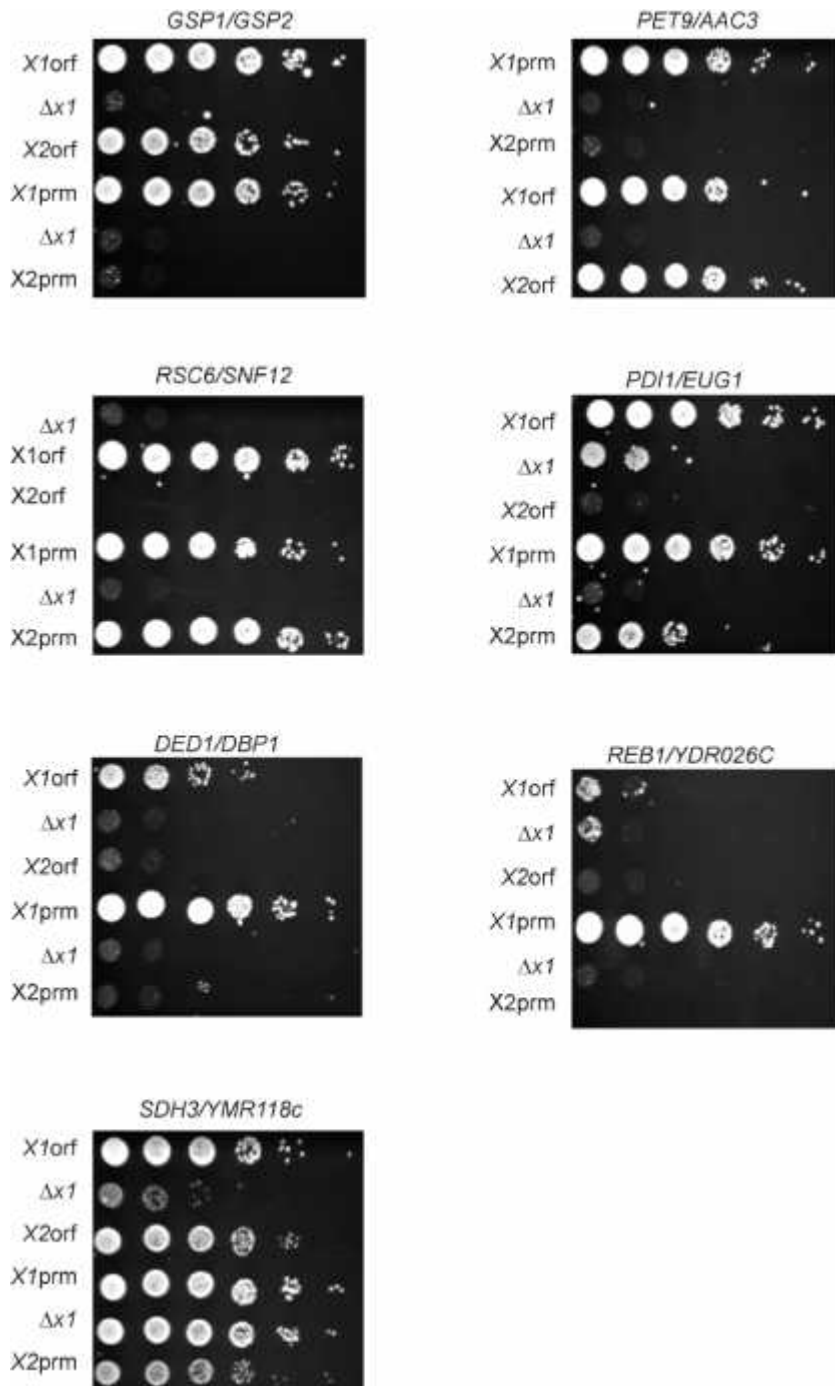


Figura 2.10 Resultado de los ensayos de viabilidad de la complementación genética de parejas parálogas con divergencia funcional extrema. Mostramos ensayos de goteo del ensayo de complementación genética de siete parejas de parálogos con al menos un integrante esencial. Las filas muestran las cinco diferentes construcciones genéticas para cada pareja de genes. Las filas de derecha a izquierda son diluciones 1:10 de cultivos saturados (1×10^6 células) de las cepas indicadas. Todas las cepas se cultivaron en medio YPAD.

Solo dos parejas de parálogos (*GSP1/2* y *PET9/AAC3*) mostraron divergencia funcional exclusivamente en el lado del promotor, lo cual, es esperado dada la alta similitud en secuencia codificante entre los parálogos de ambas parejas (ver valores *Ka* **Figure 4.10**). De hecho, ha sido reportado que la sobreexpresión de *AAC3*, que se expresa principalmente en condiciones anaerobias, puede rescatar la viabilidad en mutantes *pet9* (Kolarov, Kolarova, & Nelson, 1990). Mientras que está reportado que *GSP1* se expresa a 10 veces más que *GSP2* con glucosa como fuente de carbono (Belhumeur et al., 1993). Ambas observaciones experimentales corroboran que la pérdida del traslape funcional en estas parejas es provocada muy posiblemente por *subfuncionalización* regulatoria, aun cuando ambos parálogos realizan la misma función bioquímica, se expresan en contextos diferentes. Por el contrario, en el caso de la pareja *RSC6/SNF2* muestra una divergencia solamente del lado de la proteína, lo cual concuerda con el alto grado de divergencia en sus secuencias de aminoácidos, mientras que debido a que son genes que actúan en vías similares es posible que sus perfiles de expresión sean similares.

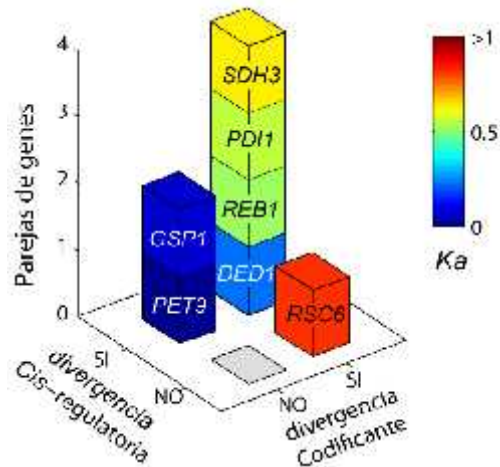


Figura 2.11 Comparación directa de la divergencia fenotípica de la región codificante y región *cis*-regulatoria de parejas parálogos con divergencia funcional extrema. El diagrama representa los resultados de viabilidad del ensayo de complementación genética. Si la complementación con el promotor u ORF parálogo es viable consideramos que no hay divergencia funcional en dicha región. Por otra parte, cuando la construcción quimérica no es viable decimos que hay una alta divergencia funcional. Indicamos el *Ka* entre las proteínas de cada pareja paróloga con un gradiente de color.

2.4.4 Nuestro ensayo con promotores regulables indica que la dosis génica de los parálogos estudiados afecta fuertemente la adecuación

El alto grado de complementación de la región regulatoria nos pareció inquietante dado el número trabajos que reportan típicamente una rápida divergencia regulatoria después de la duplicación. Una posible explicación es que la dosis génica de las parejas de duplicados seleccionadas no esté relacionada directamente con su contribución a la adecuación. Para probar esta hipótesis decidimos diseñar un sistema experimental para develar la relación existente entre la adecuación y la dosis de los genes previamente estudiados.

La estrategia usada consiste en cuantificar la adecuación relativa de cepas en las que el gen de interés es expresado en diferentes niveles, mediante el uso del promotor regulable *tetO₇*. Para encontrar el rango de concentraciones ideal del represor del sistema de expresión, expresamos la proteína fluorescente GFP bajo el promotor *tetO₇*. Después, cuantificamos

GFP a diferentes concentraciones de doxiciclina, el represor del sistema. Encontramos las concentraciones de represor que provocan mayor variación en la expresión de GFP, ver **Figura 2.12-H**, usamos este rango de concentraciones para nuestros experimentos posteriores con este sistema.

Se generaron cepas con el sistema *tetO7* para cada uno de los genes seleccionados (**Tabla 2.2**). A continuación, cuantificamos la adecuación relativa de éstas cepas en diferentes concentraciones del represor. Los resultados de estos ensayos se muestran en la **Figura 2.12-A-G**. Como es de esperarse, los parálogos cuyas mutantes de pérdida de función no tienen fenotipo (*X2*) y no presentan cambios significativos en la adecuación cuando variamos su expresión. En contraste, seis de los siete parálogos que contribuyen a la adecuación de manera significativa (*X1*) exhiben una relación cercana entre adecuación y nivel de expresión. La tendencia en cuatro casos, muestra que el nivel más alto de expresión (cuando tenemos la dosis más baja del represor) es también el punto más alto en la adecuación, lo que indica que estos genes son necesarios en dosis altas para conservar el fenotipo silvestre.

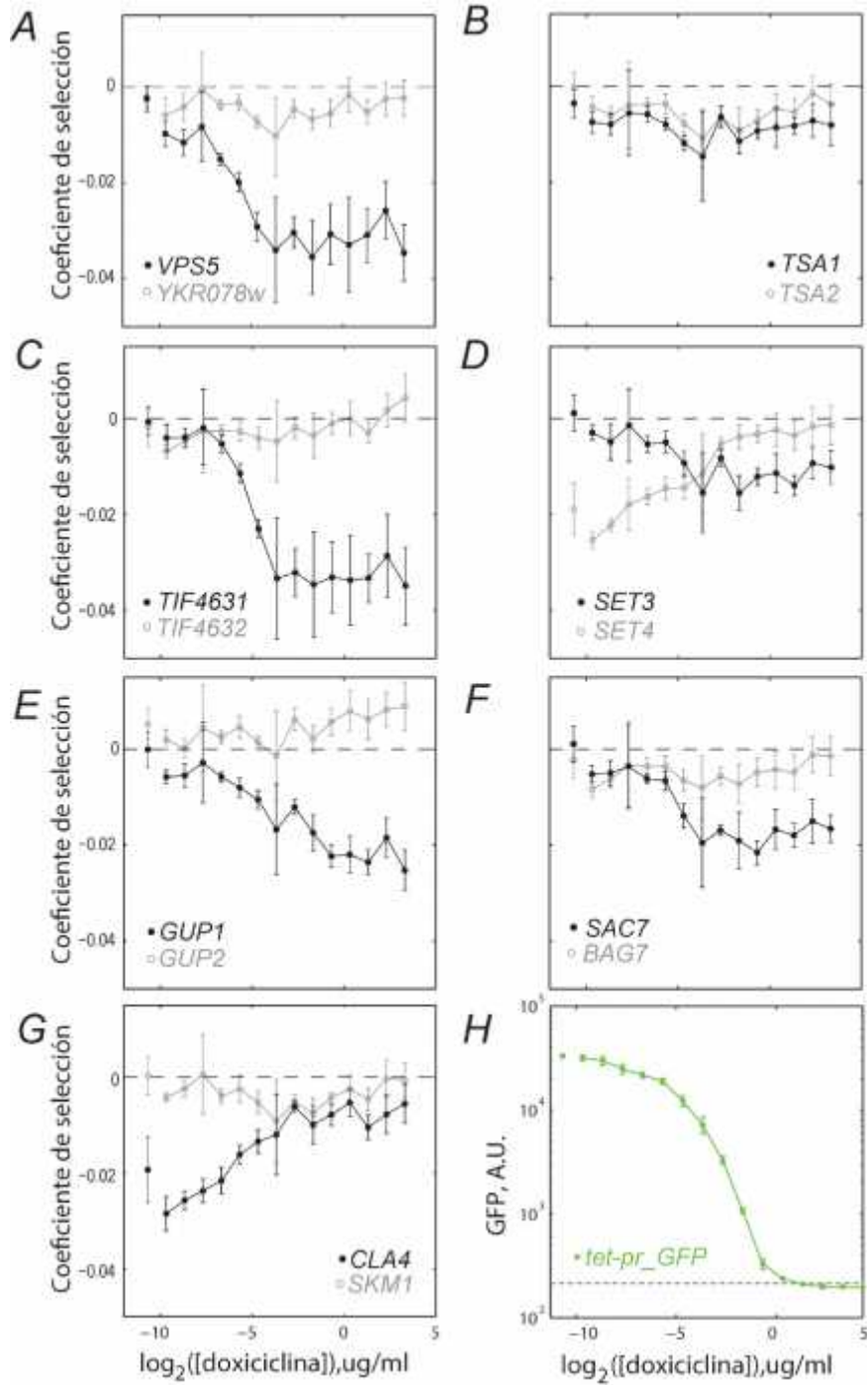


Figura 2.12 Relación entre dosis génica y adecuación. (A-G) Cepas con la secuencia codificante del gen señalado expresado bajo el promotor regulable *tetO7*. Determinamos la adecuación de dichas cepas a diferentes niveles de expresión, usando medio de competencia con diferentes concentraciones de doxiciclina. (H) La cantidad de GFP (unidades arbitrarias (A.U.)), en función de la concentración de doxiciclina.

En dos de los casos (*SET3/SET4* y *CLA4/SKMI*), ambas parejas con funciones regulatorias importantes, la relación es un poco más compleja; a niveles altos de expresión hay una reducción importante en la adecuación, mientras que el nivel más alto de adecuación es alcanzado cuando se tiene una expresión media o baja.

2.5 DISCUSIÓN

En este capítulo estudiamos la divergencia funcional de genes onhólogos mediante una estrategia experimental de alta resolución. Primero, evaluamos la contribución a la adecuación de más de 700 genes duplicados vía WGD. Encontramos que el 90% de los genes estudiados son dispensables para el crecimiento, lo que sugiere que la mayoría de los parálogos WGD conserva suficiente traslape funcional para poder compensar la ausencia de su pareja. Por otro lado, una pequeña proporción de parejas de parálogos (10%) presenta divergencia funcional notable: solo uno de los parálogos es indispensable para el crecimiento lo que revela que su parálogo ha perdido la capacidad de compensar funcionalmente su ausencia.

Tomamos una muestra de 12 parejas de genes parálogos con divergencia funcional notable para estudiar más a fondo la naturaleza de este fenómeno. Con esta finalidad, diseñamos una estrategia experimental basada en complementación genética para el estudio de la divergencia funcional de las diferentes regiones génicas. Dicho método, nos permitió examinar el papel relativo de las regiones *cis*-regulatorias y las regiones codificantes en la divergencia fenotípica de los genes duplicados mediante el cálculo de un índice de divergencia funcional.

Luego, se evaluó la divergencia funcional de siete parejas de parálogos con divergencia significativa. Se encontró que la compensación funcional es altamente dependiente de la condición, lo que denota la plasticidad de las funciones génicas que afectan los fenotipos de maneras distintas dependiendo del ambiente. Existe evidencia de divergencia fenotípica tanto en la región codificante como en la región *cis*-regulatoria para todas las parejas en al menos una de las condiciones ensayadas. Sin embargo, cuando contrastamos directamente los índices de divergencia de cada región, la divergencia fenotípica esta sesgada hacia la región codificante. Lo que sugiere que los cambios en la proteína afectan más la adecuación comparados con los cambios en el promotor.

Las parejas de parálogos estudiadas hasta este punto tienen efectos fenotípicos moderados por lo que es posible que aun conserven cierto traslape funcional. Para muestrear parejas con divergencia funcional total, se tomó un set de 8 parejas de genes en las que uno de los parálogos es esencial mientras que el otro parálogo es dispensable. Como es esperado la mayor parte de estas parejas (5/8) presentan divergencia funcional total en ambas regiones génicas. Dos parejas presentaron divergencia funcional en la región *cis*-regulatoria pero una alta conservación en la secuencia. Lo que denota una diversificación de sus funciones regulatorias y una conservación su función bioquímica.

El hecho de que los promotores complementaran mejor que las proteínas nos inquietó debido a múltiples reportes de la rápida divergencia regulatoria de genes parálogos después de la duplicación. Una posibilidad es que la relación entre la contribución a la adecuación y el nivel de expresión no es proporcional en el caso de nuestros parálogos. Por lo que, las proteínas expresadas a niveles variables pueden compensar igualmente la falta de su parólogo. Con esta interrogante en mente, describimos la relación de la adecuación con la expresión de

todas nuestras parejas parálogas. Sorprendentemente, encontramos que para la mayoría de los genes la relación entre dosis génica y adecuación es muy cercana, solo un rango muy limitado de expresión complementa el promotor nativo. Lo cual sugiere fuertemente que hay una complementación genuina entre los promotores parálogos, debido muy posiblemente a la conservación de elementos regulatorios.

DISCUSIÓN GENERAL

Este trabajo es una aproximación integral al estudio de la duplicación génica en *S. cerevisiae*. Aprovechamos la facilidad de generar nuevas estrategias sistemáticas tanto bioinformáticas y experimentales para abordar el estudio de preguntas abiertas de la vida evolutiva de genes duplicados en su camino a la generación de funciones novedosas y con ello generar diversidad fenotípica.

En la primera parte de este trabajo calculamos las tasas de evolución antes y después de la WGD analizando secuencias de 23 especies ascomicetos. Nuestros resultados confirman que para la mayor parte de los casos analizados existe un incremento significativo de las tasas dN/dS en las filogenias de parálogos WGD. Identificamos un grupo de genes sujetos a una fuerte presión de selección antes de la WGD, el cual prácticamente en todos los casos aceleró sus tasas de evolución después de la duplicación. Esto sugiere que la duplicación genera un efecto de relajación en la selección purificadora mayormente en proteínas altamente conservadas.

Comparamos los patrones evolutivos de diferentes clases funcionales. Encontramos que las proteínas ribosomales tienden a conservar la tasa de evolución ancestral, mientras que por el contrario las proteínas regulatorias (quinasas y factores de transcripción) son muy propensas a acelerar su tasa después de la duplicación. Es importante recordar que las familias de genes regulatorios se expanden preferencialmente por WGD, lo que resulta en una ampliación de las redes regulatorias y de señalización (Corrochano et al., 2016). De manera interesante, los factores de transcripción se encuentran en el grupo de genes con una fuerte selección purificadora antes de la WGD, por lo que la duplicación génica juega un papel importante en

la diversificación de estos genes; permitiéndoles librarse de éstas fuertes restricciones funcionales. Esto sugiere que la WGD juega un papel fundamental en la ampliación del repertorio regulatorio de eucariontes y consecuentemente de su gran diversidad morfológica (Canestro, Albalat, Irimia, & Garcia-Fernandez, 2013; Holland & Garcia-Fernández, 1996).

Las proteínas que aceleran sus tasas de evolución después de la duplicación comparten ciertas características: son menos abundantes, tienen mayor contenido de desorden intrínseco y contribuyen menos a la adecuación que genes con tasas de evolución constantes. Sin embargo encontramos que dichas características se relacionan también con la clase funcional. Las proteínas ribosomales son muy abundantes, son muy ordenadas y contribuyen de manera importante a la adecuación; y se encuentran en el grupo de genes que no cambian su tasa de evolución. Por el contrario, las proteínas reguladoras tienden a ser más desordenadas y a contribuir menos a la adecuación; además evolucionan más rápido después de la duplicación. De manera que sub o sobre representación de estos grupos funcionales en el grupo de genes que cambian sus tasas de evolución da como resultado las tendencias que vemos en estas características moleculares y fenotípicas. De hecho, ya se ha sugerido que la función de la proteína es la que determina las restricciones a los cambios en su secuencia (Hirsh & Fraser, 2001).

Encontramos que un poco más de la mitad de los ortogrupos analizados exhiben tasas de evolución asimétricas entre los parálogos, lo cual coincide con otros estudios en levadura y otros organismos (Brunet et al., 2006; Conant & Wagner, 2003; Scannell & Wolfe, 2008; P. Zhang, Gu, & Li, 2003). No encontramos una asociación entre la abundancia de proteína y este patrón de evolución asimétrica.

Una limitante importante de los estudios *in silico* de tasas de evolución es que se desconoce hasta qué punto estos cambios producen fenotipos. Además, en estos análisis ignoramos por completo cambios importantes que puedan afectar la regulación de los genes. Más aún, hay muy pocos trabajos que contrasten el efecto de los cambios en secuencia y regulación a nivel fenotípico. Con esta finalidad, en la segunda parte estimamos de manera experimental la divergencia fenotípica de la región *cis*-regulatoria y codificante de un pequeño grupo de parálogos producto de la WGD. Nuestra estrategia experimental de alta resolución arrojó resultados que sugieren que la divergencia fenotípica de los genes duplicados es dirigida en conjunto por la región *cis*-regulatoria y codificante.

Existen múltiples estudios que sugieren que la regulación diverge rápidamente después de la duplicación (Gu et al., 2002; B. Papp et al., 2003; Tirosh & Barkai, 2007), sin embargo, rara vez se ha medido su impacto en la adecuación. A pesar de que en casi todos los casos estudiados hay una fuerte divergencia fenotípica del promotor dicha divergencia está asociada con una divergencia de magnitud parecida de la proteína. Para la mayoría de los casos los promotores parálogos son más intercambiables que las proteínas, lo que sugiere que aunque la región *cis*-regulatoria diverge rápidamente después de la duplicación, se mantiene cierta capacidad de compensación funcional. Esta compensación entre los promotores parálogos no es trivial, ya que un ensayo con un promotor regulable revela que no cualquier nivel de expresión restaura la pérdida del promotor endógeno. Esto concuerda con un escenario de en el que el promotor parálogo responde ante la ausencia de su contraparte, donde ciertos mecanismos permiten garantizar la expresión de un gen en caso de necesidad y mantener una robustez en el sistema biológico (Diss, Ascencio, DeLuna, & Landry, 2014).

CONCLUSIONES

En la primera parte de este proyecto medimos la acción de la selección natural en diferentes puntos de la historia evolutiva de parálogos WGD, llegando a las siguientes conclusiones:

- J Las tasas de evolución de las secuencias codificantes de genes parálogos aumentan después de la WGD en la mayor parte de los ortogrupos estudiados, lo que concuerda con estudios similares.
- J Notamos que los genes más propensos a aumentar su tasa de evolución estuvieron sujetos a una presión de selección muy fuerte previa a la WGD, lo que denota la importancia de la duplicación en el proceso de exploración de funciones novedosas.
- J Cuando comparamos las tasas en diferentes puntos de la filogenia, observamos que las ramas que divergen inmediatamente después de la duplicación experimentan una aceleración súbita en sus tasas de evolución, posteriormente, en los clados modernos esta tasa disminuye y se estabiliza, pero no regresa al mismo nivel que las secuencias ancestrales.
- J Las propiedades como abundancia, la contribución a la adecuación y el desorden intrínseco de una proteína están correlacionados de manera significativa con el cambio en las tasas de evolución después de la duplicación.
- J Sin embargo, cuando examinamos las categorías funcionales más representativas de nuestro set de datos (metabolismo, regulación y ribosomas), encontramos que dichas propiedades están altamente relacionadas con estas funciones, por lo que su correlación con el cambio en las tasas de evolución es un producto de la categoría funcional.

Luego, examinamos el papel relativo de las regiones *cis*-regulatorias y las regiones codificantes en la divergencia fenotípica de los genes duplicados mediante una estrategia experimental de alta resolución basada en complementación genética:

-) La mayoría de los ohnólogos son capaces de compensarse funcionalmente en condiciones estándar de laboratorio. Sin embargo, encontramos que aproximadamente 10% de éstas parejas de genes presentan divergencia fenotípica en forma asimétrica; mientras solo uno de los parálogos compensa la ausencia de su pareja.
-) Observamos evidencia experimental de divergencia funcional tanto en promotor como en proteína de nuestro grupo de genes duplicados.
-) Sin embargo, los cambios en secuencia de aminoácidos de las proteínas contribuyen mayormente a la divergencia funcional de éstos parálogos.
-) Nuestro análisis de promotores regulables sugiere que la adecuación está fuertemente correlacionada con la expresión de nuestro grupo de parálogos. Mientras que los promotores parálogos son más intercambiables debido a que conservan cierta capacidad de compensación.

PREGUNTAS ABIERTAS

Componentes responsables de la divergencia codificante de genes duplicados.

En el segundo capítulo observamos que varias de nuestras parejas de genes presentan divergencia funcional notable en sus proteínas. Sin embargo, dicha divergencia no siempre se ve reflejada en la tasa de mutaciones no sinónimas por lo que es de esperarse que los cambios que modifican la función sean puntuales, muy probablemente en sitios clave como sitios activos o residuos de aminoácidos relevantes para una correcta conformación estructural de la proteína. El siguiente paso natural del proyecto es describir la naturaleza de dichos cambios. Una estrategia es tomar una proteína paróloga y generar diversas variantes mediante mutagénesis aleatoria. Generar mutantes de levadura que expresen las diferentes variantes y cuantificar su adecuación. Por último aquellas variantes que presenten pérdida de adecuación pueden secuenciarse para determinar que regiones de la proteína son responsables por dicha pérdida.

Una segunda estrategia, es determinar *in silico* la tasa de evolución individual de cada codón de una proteína paróloga, seleccionar aquellos codones en los que haya una relajación en la presión de selección o exista selección positiva, especialmente aquellos sitios que difieran entre parálogos y generar mutantes en dichos sitios mediante mutagénesis sitio dirigida. Por último, cuantificar la adecuación de dichas mutantes para evaluar la importancia funcional de dicho sitio.

Estudiar cambios en secuencias *cis*-regulatorias de genes parálogos

En la primera parte de este proyecto nos dedicamos a estudiar los cambios en la secuencia codificante de genes parálogos. Pero, como observamos en el segundo capítulo las regiones

regulatorias contribuyen también de manera importante a la divergencia funcional de genes duplicados y por tanto a la innovación funcional. Una posibilidad para estudiar los cambios en dichas regiones es detectar sitios de unión a factores de transcripción y como se conservan a lo largo de los ortogrupos y especialmente entre parálogos con diferentes grados de divergencia funcional o de secuencia.

El efecto fenotípico de la duplicación en escala sencilla.

En este trabajo estudiamos varias etapas en la vida evolutiva de genes parálogos, centrándonos en genes que se han conservado durante largos tiempos evolutivos. Se sabe, sin embargo, que el destino de la mayoría de las duplicaciones es perderse, mientras que la mayoría de los genes ni si quiera llegan a fijarse en la población. Por otro lado, como discutimos en la introducción, hay grupos de genes que tienden a duplicarse constantemente o por el contrario se duplican rara vez. ¿Qué tipo de fuerzas evolutivas determinan la fijación o pérdida de un gen recién duplicado?, en particular nos preguntamos si la selección natural juega un papel relevante en la fijación de genes duplicados. Si es así esperaríamos que los genes que tienden a encontrarse en más de una copia generen un efecto benéfico al duplicarse. O por el contrario genes que nunca se duplican, acarreen defectos adecuacionales cuando se encuentran en más de una copia. En uno de nuestros proyectos actuales abordamos esta pregunta mediante una estrategia experimental: duplicamos individualmente un set de genes en levadura, con su propia secuencia *cis*-regulatoria, y evaluamos el efecto fenotípico de dicha modificación.

Nuestros resultados preliminares para la duplicación experimental de más de 900 genes esenciales de levadura, indican que la duplicación génica puede acarrear tanto efectos benéficos como deletéreos para una buena proporción de los genes ensayados (9 y 12%

respectivamente). El siguiente paso es estudiar las propiedades y mecanismos moleculares asociados con estos efectos fenotípicos.

REFERENCIAS

- Aasland, R., Gibson, T. J., & Stewart, A. F. (1995). The PHD finger: implications for chromatin-mediated transcriptional regulation. *Trends Biochem Sci*, 20(2), 56-59.
- Adams, K. L., & Wendel, J. F. (2005). Polyploidy and genome evolution in plants. *Curr Opin Plant Biol*, 8(2), 135-141. doi: 10.1016/j.pbi.2005.01.001
- Baudin, A., Ozier-Kalogeropoulos, O., Denouel, A., Lacroute, F., & Cullin, C. (1993). A simple and efficient method for direct gene deletion in *Saccharomyces cerevisiae*. *Nucleic Acids Res*, 21(14), 3329-3330.
- Belhumeur, P., Lee, A., Tam, R., DiPaolo, T., Fortin, N., & Clark, M. W. (1993). GSP1 and GSP2, genetic suppressors of the *prp20-1* mutant in *Saccharomyces cerevisiae*: GTP-binding proteins involved in the maintenance of nuclear organization. *Mol Cell Biol*, 13(4), 2152-2161.
- Breker, M., Gymrek, M., & Schuldiner, M. (2013). A novel single-cell screening platform reveals proteome plasticity during yeast stress responses. *The Journal of cell biology*, 200(6), 839-850. doi: 10.1083/jcb.201301120
- Breslow, D. K., Cameron, D. M., Collins, S. R., Schuldiner, M., Stewart-Ornstein, J., Newman, H. W., . . . Weissman, J. S. (2008). A comprehensive strategy enabling high-resolution functional analysis of the yeast genome. *Nat Methods*, 5(8), 711-718. doi: 10.1038/nmeth.1234
- Brown, C. J., Takayama, S., Campen, A. M., Vise, P., Marshall, T. W., Oldfield, C. J., . . . Dunker, A. K. (2002). Evolutionary rate heterogeneity in proteins with long disordered regions. *J Mol Evol*, 55(1), 104-110. doi: 10.1007/s00239-001-2309-6
- Brunet, F. G., Roest Crollius, H., Paris, M., Aury, J. M., Gibert, P., Jaillon, O., . . . Robinson-Rechavi, M. (2006). Gene loss and evolutionary rates following whole-genome duplication in teleost fishes. *Mol Biol Evol*, 23(9), 1808-1816. doi: 10.1093/molbev/msl049
- Byrne, K. P., & Wolfe, K. H. (2007). Consistent patterns of rate asymmetry and gene loss indicate widespread neofunctionalization of yeast genes after whole-genome duplication. *Genetics*, 175(3), 1341-1350. doi: 10.1534/genetics.106.066951
- Canestro, C., Albalat, R., Irimia, M., & Garcia-Fernandez, J. (2013). Impact of gene gains, losses and duplication modes on the origin and diversification of vertebrates. *Semin Cell Dev Biol*, 24(2), 83-94. doi: 10.1016/j.semcdb.2012.12.008
- Cavaliere, D., McGovern, P. E., Hartl, D. L., Mortimer, R., & Polsinelli, M. (2003). Evidence for *S. cerevisiae* fermentation in ancient wine. *J Mol Evol*, 57 Suppl 1, S226-232. doi: 10.1007/s00239-003-0031-2
- Conant, G. C., & Wagner, A. (2003). Asymmetric sequence divergence of duplicate genes. *Genome research*, 13(9), 2052-2058. doi: 10.1101/gr.1252603
- Conant, G. C., & Wolfe, K. H. (2008). Turning a hobby into a job: how duplicated genes find new functions. *Nat Rev Genet*, 9(12), 938-950. doi: 10.1038/nrg2482
- Corrochano, L. M., Kuo, A., Marcet-Houben, M., Polaino, S., Salamov, A., Villalobos-Escobedo, J. M., . . . Grigoriev, I. V. (2016). Expansion of Signal Transduction Pathways in Fungi by

- Extensive Genome Duplication. *Curr Biol*, 26(12), 1577-1584. doi: 10.1016/j.cub.2016.04.038
- Choi, S. S., & Hannenhalli, S. (2013). Three independent determinants of protein evolutionary rate. *Journal of molecular evolution*, 76(3), 98-111. doi: 10.1007/s00239-013-9543-6
- Davis, J. C., & Petrov, D. A. (2004). Preferential duplication of conserved proteins in eukaryotic genomes. *PLoS Biol*, 2(3), E55. doi: 10.1371/journal.pbio.0020055
- Dean, E. J., Davis, J. C., Davis, R. W., & Petrov, D. A. (2008). Pervasive and persistent redundancy among duplicated genes in yeast. *PLoS Genet*, 4(7), e1000113. doi: 10.1371/journal.pgen.1000113
- Dehal, P., & Boore, J. L. (2005). Two rounds of whole genome duplication in the ancestral vertebrate. *PLoS Biol*, 3(10), e314.
- DeLuna, A., Vetsigian, K., Shoresh, N., Hegreness, M., Colon-Gonzalez, M., Chao, S., & Kishony, R. (2008). Exposing the fitness contribution of duplicated genes. *Nat Genet*, 40(5), 676-681. doi: 10.1038/ng.123
- Diss, G., Ascencio, D., DeLuna, A., & Landry, C. R. (2014). Molecular mechanisms of paralogous compensation and the robustness of cellular networks. *J Exp Zool B Mol Dev Evol*, 322(7), 488-499. doi: 10.1002/jez.b.22555
- Drummond, D. A., Bloom, J. D., Adami, C., Wilke, C. O., & Arnold, F. H. (2005). Why highly expressed proteins evolve slowly. *Proc Natl Acad Sci U S A*, 102(40), 14338-14343. doi: 10.1073/pnas.0504070102
- Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research*, 32(5), 1792-1797. doi: 10.1093/nar/gkh340
- Force, A., Lynch, M., Pickett, F. B., Amores, A., Yan, Y. L., & Postlethwait, J. (1999). Preservation of duplicate genes by complementary, degenerative mutations. *Genetics*, 151(4), 1531-1545.
- Gagnon-Arsenault, I., Marois Blanchet, F. C., Rochette, S., Diss, G., Dube, A. K., & Landry, C. R. (2013). Transcriptional divergence plays a role in the rewiring of protein interaction networks after gene duplication. *J Proteomics*, 81, 112-125. doi: 10.1016/j.jprot.2012.09.038
- Ghaemmaghami, S., Huh, W. K., Bower, K., Howson, R. W., Belle, A., Dephoure, N., . . . Weissman, J. S. (2003). Global analysis of protein expression in yeast. *Nature*, 425(6959), 737-741. doi: 10.1038/nature02046
- Giaever, G., Chu, A. M., Ni, L., Connelly, C., Riles, L., Veronneau, S., . . . Johnston, M. (2002). Functional profiling of the *Saccharomyces cerevisiae* genome. *Nature*, 418(6896), 387-391. doi: 10.1038/nature00935
- Goffeau, A., Barrell, B. G., Bussey, H., Davis, R. W., Dujon, B., Feldmann, H., . . . Oliver, S. G. (1996). Life with 6000 genes. *Science*, 274(5287), 546, 563-547.
- Gout, J.-F., Kahn, D., Duret, L., & Consortium, P. P.-G. (2010). The relationship among gene expression, the evolution of gene dosage, and the rate of protein evolution.
- Gu, Z., Nicolae, D., Lu, H. H., & Li, W. H. (2002). Rapid divergence in expression between duplicate genes inferred from microarray data. *Trends Genet*, 18(12), 609-613.
- Guan, Y., Dunham, M. J., & Troyanskaya, O. G. (2007). Functional analysis of gene duplications in *Saccharomyces cerevisiae*. *Genetics*, 175(2), 933-943. doi: 10.1534/genetics.106.064329

- Hahn, M. W. (2009). Distinguishing among evolutionary models for the maintenance of gene duplicates. *J Hered*, *100*(5), 605-617. doi: 10.1093/jhered/esp047
- Hakes, L., Pinney, J. W., Lovell, S. C., Oliver, S. G., & Robertson, D. L. (2007). All duplicates are not equal: the difference between small-scale and genome duplication. *Genome Biol*, *8*(10), R209. doi: 10.1186/gb-2007-8-10-r209
- Halligan, D. L., Kousathanas, A., Ness, R. W., Harr, B., Eory, L., Keane, T. M., . . . Keightley, P. D. (2013). Contributions of protein-coding and regulatory change to adaptive molecular evolution in murid rodents. *PLoS Genet*, *9*(12), e1003995. doi: 10.1371/journal.pgen.1003995
- Herskowitz, I. (1988). Life cycle of the budding yeast *Saccharomyces cerevisiae*. *Microbiol Rev*, *52*(4), 536-553.
- Hirsh, A. E., & Fraser, H. B. (2001). Protein dispensability and rate of evolution. *Nature*, *411*(6841), 1046-1049. doi: 10.1038/35082561
- Hittinger, C. T., & Carroll, S. B. (2007). Gene duplication and the adaptive evolution of a classic genetic switch. *Nature*, *449*(7163), 677-681. doi: 10.1038/nature06151
- Ho, C. H., Magtanong, L., Barker, S. L., Gresham, D., Nishimura, S., Natarajan, P., . . . Boone, C. (2009). A molecular barcoded yeast ORF library enables mode-of-action analysis of bioactive compounds. *Nat Biotechnol*, *27*(4), 369-377. doi: 10.1038/nbt.1534
- Holland, P. W., & Garcia-Fernàndez, J. (1996). Hoxgenes and chordate evolution. *Developmental biology*, *173*(2), 382-395.
- Hughes, A. L. (1994). The evolution of functionally novel proteins after gene duplication. *Proc Biol Sci*, *256*(1346), 119-124. doi: 10.1098/rspb.1994.0058
- Hughes, T. R., Marton, M. J., Jones, A. R., Roberts, C. J., Stoughton, R., Armour, C. D., . . . Friend, S. H. (2000). Functional discovery via a compendium of expression profiles. *Cell*, *102*(1), 109-126.
- Innan, H., & Kondrashov, F. (2010). The evolution of gene duplications: classifying and distinguishing between models. *Nature Reviews Genetics*, *11*(2), 97-108.
- Iost, I., Dreyfus, M., & Linder, P. (1999). Ded1p, a DEAD-box protein required for translation initiation in *Saccharomyces cerevisiae*, is an RNA helicase. *J Biol Chem*, *274*(25), 17677-17683.
- Jones, D. T., & Ward, J. J. (2003). Prediction of disordered regions in proteins from position specific score matrices. *Proteins*, *53 Suppl 6*, 573-578. doi: 10.1002/prot.10528
- Kellis, M., Birren, B. W., & Lander, E. S. (2004). Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*. *Nature*, *428*(6983), 617-624. doi: 10.1038/nature02424
- Kolarov, J., Kolarova, N., & Nelson, N. (1990). A third ADP/ATP translocator gene in yeast. *J Biol Chem*, *265*(21), 12711-12716.
- Kondrashov, F. A., Rogozin, I. B., Wolf, Y. I., & Koonin, E. V. (2002). Selection in the evolution of gene duplications. *Genome Biol*, *3*(2), RESEARCH0008.
- Krakauer, D. C., & Nowak, M. A. (1999). Evolutionary preservation of redundant duplicated genes. *Semin Cell Dev Biol*, *10*(5), 555-559. doi: 10.1006/scdb.1999.0337

- Landry, C. R., Levy, E. D., & Michnick, S. W. (2009). Weak functional constraints on phosphoproteomes. *Trends Genet*, 25(5), 193-197. doi: 10.1016/j.tig.2009.03.003
- Li, L., Huang, Y., Xia, X., & Sun, Z. (2006). Preferential duplication in the sparse part of yeast protein interaction network. *Molecular biology and evolution*, 23(12), 2467-2473.
- Lynch, M., & Conery, J. S. (2000). The evolutionary fate and consequences of duplicate genes. *Science*, 290(5494), 1151-1155.
- Maere, S., De Bodt, S., Raes, J., Casneuf, T., Van Montagu, M., Kuiper, M., & Van de Peer, Y. (2005). Modeling gene and genome duplications in eukaryotes. *Proc Natl Acad Sci U S A*, 102(15), 5454-5459. doi: 10.1073/pnas.0501102102
- Nowak, M. A., Boerlijst, M. C., Cooke, J., & Smith, J. M. (1997). Evolution of genetic redundancy. *Nature*, 388(6638), 167-171. doi: 10.1038/40618
- Ohno, S. (1970). *Evolution by gene duplication*: Springer Science & Business Media.
- Otto, S. P. (2007). The evolutionary consequences of polyploidy. *Cell*, 131(3), 452-462. doi: 10.1016/j.cell.2007.10.022
- Pal, C., Papp, B., & Hurst, L. D. (2001). Highly expressed genes in yeast evolve slowly. *Genetics*, 158(2), 927-931.
- Papp, B., Pal, C., & Hurst, L. D. (2003). Dosage sensitivity and the evolution of gene families in yeast. *Nature*, 424(6945), 194-197.
- Papp, B., Pal, C., & Hurst, L. D. (2003). Evolution of cis-regulatory elements in duplicated genes of yeast. *Trends Genet*, 19(8), 417-422. doi: 10.1016/s0168-9525(03)00174-4
- Pegueroles, C., Laurie, S., & Albà, M. M. (2013). Accelerated evolution after gene duplication: a time-dependent process affecting just one copy. *Molecular biology and evolution*, mst083.
- Ruderfer, D. M., Pratt, S. C., Seidel, H. S., & Kruglyak, L. (2006). Population genomic analysis of outcrossing and recombination in yeast. *Nat Genet*, 38(9), 1077-1081. doi: 10.1038/ng1859
- Samuel, D. (1996). Investigation of Ancient Egyptian Baking and Brewing Methods by Correlative Microscopy. *Science*, 273(5274), 488-490.
- Scannell, D. R., & Wolfe, K. H. (2008). A burst of protein sequence evolution and a prolonged period of asymmetric evolution follow gene duplication in yeast. *Genome Res*, 18(1), 137-147. doi: 10.1101/gr.6341207
- Seoighe, C., & Wolfe, K. H. (1999). Yeast genome evolution in the post-genome era. *Current opinion in microbiology*, 2(5), 548-554.
- Sheff, M. A., & Thorn, K. S. (2004). Optimized cassettes for fluorescent protein tagging in *Saccharomyces cerevisiae*. *Yeast*, 21(8), 661-670. doi: 10.1002/yea.1130
- Suyama, M., Torrents, D., & Bork, P. (2006). PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res*, 34(Web Server issue), W609-612. doi: 10.1093/nar/gkl315
- Teichmann, S. A., & Babu, M. M. (2004). Gene regulatory network growth by duplication. *Nat Genet*, 36(5), 492-496. doi: 10.1038/ng1340
- Tirosh, I., & Barkai, N. (2007). Comparative analysis indicates regulatory neofunctionalization of yeast duplicates. *Genome Biol*, 8(4), R50. doi: 10.1186/gb-2007-8-4-r50

- Tong, A. H., & Boone, C. (2006). Synthetic genetic array analysis in *Saccharomyces cerevisiae*. *Methods Mol Biol*, 313, 171-192.
- Veitia, R. A., Bottani, S., & Birchler, J. A. (2008). Cellular reactions to gene dosage imbalance: genomic, transcriptomic and proteomic effects. *Trends in Genetics*, 24(8), 390-397.
- Veitia, R. A., & Potier, M. C. (2015). Gene dosage imbalances: action, reaction, and models. *Trends Biochem Sci*, 40(6), 309-317. doi: 10.1016/j.tibs.2015.03.011
- Wach, A., Brachat, A., Pohlmann, R., & Philippsen, P. (1994). New heterologous modules for classical or PCR-based gene disruptions in *Saccharomyces cerevisiae*. *Yeast*, 10(13), 1793-1808.
- Wapinski, I., Pfeffer, A., Friedman, N., & Regev, A. (2007). Automatic genome-wide reconstruction of phylogenetic gene trees. *Bioinformatics*, 23(13), i549-558. doi: 10.1093/bioinformatics/btm193
- Wapinski, I., Pfeffer, A., Friedman, N., & Regev, A. (2007). Natural history and evolutionary principles of gene duplication in fungi. *Nature*, 449(7158), 54-61. doi: 10.1038/nature06107
- Wolfe, K. H., & Shields, D. C. (1997). Molecular evidence for an ancient duplication of the entire yeast genome. *Nature*, 387(6634), 708-712.
- Yang, Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. *Molecular biology and evolution*, 24(8), 1586-1591.
- Yen, K., Gitsham, P., Wishart, J., Oliver, S. G., & Zhang, N. (2003). An improved tetO promoter replacement system for regulating the expression of yeast genes. *Yeast*, 20(15), 1255-1262. doi: 10.1002/yea.1037
- Zhang, J. (2003). Evolution by gene duplication: an update. *Trends in Ecology & Evolution*, 18(6), 292-298.
- Zhang, J., & Yang, J. R. (2015). Determinants of the rate of protein sequence evolution. *Nat Rev Genet*, 16(7), 409-420. doi: 10.1038/nrg3950
- Zhang, P., Gu, Z., & Li, W. H. (2003). Different evolutionary patterns between young duplicate genes in the human genome. *Genome Biol*, 4(9), R56. doi: 10.1186/gb-2003-4-9-r56

ANEXO A

Ejemplo de archivo control de CodeML

```
seqfile = ../Ohno/alignments/YCL024W.e1.aln      * sequence data
filename
treefile = ../Ohno/R2/YCL024W.tre              * tree structure file
name
outfile = ../Ohno/R2-lnL/YCL024W.R2.txt        * main result file
name

noisy      =3      * 0,1,2,3,9: how much rubbish on the screen

verbose    =1      * 1:detailed output

runmode    =0      * 0:user defined tree

seqtype    =1      * 1:codons

CodonFreq  =2      * 0:equal, 1:F1X4, 2:F3X4, 3:F61

model = 2   *0:one omega ratio for all branches
           *1:separate omega for each branch
           *2:user specified dN/dS ratios for branches

NSsites    =0      *

icode      =0      * 0:universal code

fix_kappa  =0      * 1:kappa fixed, 0:kappa to be estimated
kappa     =2      * initial or fixed kappa

fix_omega  =0      * 1:omega fixed, 0:omega to be estimated
omega     =1      * initial omega

fix_blengt = 0    * 0: ignore, -1: random, 1: initial, 2: fixed

      *comments:
      *H0 in Table 3: model = 0
      *H1 in Table 3: model = 2
      *H2 in Table 3: model = 2
      *H3 in Table 3: model = 2
```

ANEXO B

Resultados de los análisis de CodeML

ORF A	Gen A	ORF B	Gen B	ω_0	ω_1	LRT	ω_1	ω_2	LRT	ω_3	ω_4	LRT
				test1	test2	1	test 2	test2	2	test 3	test 3	3
YAL015C	NTG1	YOL043C	NTG2	0.006	0.12377	1	0.04	0.12681	0	0.1676	0.08838	1
YAL017W	PSK1	YOL045W	PSK2	0.006	0.06549	1	212.9	0.06542	1	0.0635	0.06869	0
YAL023C	PMT2	YOR321W	PMT3	NA	NA	NA	NA	NA	NA	NA	NA	NA
YAL028W	FRT2	YOR324C	FRT1	0.012	0.14961	1	999	0.14937	0	0.1648	0.13431	0
YAL029C	MYO4	YOR326W	MYO2	0.04	0.06921	1	0.544	0.06781	1	0.1166	0.041	1
YAL030W	SNC1	YOR327C	SNC2	0.021	0.03644	1	999	0.03221	1	0.0294	0.03389	0
YAL034C	FUN19	YOR338W	YOR338W	0.08	0.12255	1	999	0.11963	0	0.0957	0.14597	1
YAL037W	YAL037W	YOR342C	YOR342C	0.079	0.1445	1	0.009	0.15021	1	0.121	0.20993	1
YAL038W	CDC19	YOR347C	PYK2	0.12	0.06507	1	0.159	0.06018	1	0.0391	0.12883	1
YAL051W	OAF1	YOR363C	PIP2	0.008	0.06915	1	45.71	0.06905	1	0.0669	0.07136	0
YAL053W	FLC2	YOR365C	YOR365C	0.083	0.07598	0	0.024	0.07611	0	0.1222	0.04432	1
YAL056W	GPB2	YOR371C	GPB1	0.008	0.11035	1	999	0.10947	0	0.0981	0.12136	1
YAL062W	GDH3	YOR375C	GDH1	0.135	0.06441	1	0.491	0.05522	1	0.059	0.04278	0
YAR002C-A	ERP1	YGL002W	ERP6	0.065	0.10105	1	0.531	0.09787	0	0.096	0.10022	0
YBL005W	PDR3	YGL013C	PDR1	0.021	0.10681	1	999	0.10594	0	0.0939	0.1187	1
YBL009W	ALK2	YGL021W	ALK1	0.01	0.12249	1	0.126	0.12248	0	0.1572	0.09379	1
YBL030C	PET9	YBR085W	AAC3	NA	NA	NA	NA	NA	NA	NA	NA	NA
YBL039C	URA7	YJR103W	URA8	0.037	0.04918	1	0.051	0.04899	0	0.0442	0.05794	0
YBL043W	ECM13	YJR115W	YJR115W	0.014	0.14654	1	0.306	0.14537	0	0.0417	0.15679	1
YBL054W	TOD6	YER088C	DOT6	0.067	0.15591	1	346	0.15555	1	0.126	0.18267	1
YBL056W	PTC3	YER089C	PTC2	0.024	0.05058	1	999	0.04944	1	0.0388	0.05956	1
YBL059W	YBL059W	YER093C-A	AIM11	0.011	0.13241	1	999	0.13168	0	0.1405	0.12356	0
YBL061C	SKT5	YER096W	SHC1	0.073	0.1298	1	0.179	0.12828	0	0.1912	0.10471	1
YBL067C	UBP13	YER098W	UBP9	0.005	0.12357	1	70.65	0.12186	0	0.1163	0.12702	0
YBL068W	PRS4	YER099C	PRS2	0.021	0.02092	0	999	0.02021	1	0.0205	0.01994	0
YBL069W	AST1	YER101C	AST2	0.028	0.06031	1	0.073	0.05985	0	0.0483	0.06761	1
YBL072C	RPS8A	YER102W	RPS8B	0.115	0.09074	0	1E-04	0.09074	0	0.073	0.16317	1
YBL075C	SSA3	YER103W	SSA4	0.059	0.04875	1	999	0.04677	1	0.0505	0.04112	1
YBL079W	NUP170	YER105C	NUP157	0.028	0.10578	1	0.06	0.10754	1	0.0671	0.1488	1
YBL085W	BOI1	YER114C	BOI2	0.007	0.10597	1	88.47	0.10505	1	0.1118	0.0987	0
YBL087C	RPL23A	YER117W	RPL23B	0.049	0.02909	0	1E-04	0.02993	0	NA	NA	NA
YBL089W	AVT5	YER119C	AVT6	0.05	0.08087	1	0.215	0.08073	0	NA	NA	NA
YBL091C-A	SCS22	YER120W	SCS2	0.107	0.19873	1	0.186	0.19911	0	0.2267	0.17736	0
YBL101C	ECM21	YPR030W	CSR2	0.007	0.11334	1	825.1	0.11247	0	0.0912	0.13673	1
YBL106C	SRO77	YPR032W	SRO7	0.029	0.09326	1	0.078	0.09376	0	0.0996	0.08869	0
YBR001C	NTH2	YDR001C	NTH1	0.02	0.0321	1	999	0.03191	1	0.042	0.02137	1
YBR005W	RCR1	YDR003W	RCR2	0.162	0.15391	0	0.216	0.15309	0	0.1213	0.19003	1
YBR014C	GRX7	YDL010W	GRX6	0.078	0.17457	1	0.043	0.18181	0	0.2182	0.14558	1
YBR020W	GAL1	YDR009W	GAL3	NA	NA	NA	NA	NA	NA	NA	NA	NA
YBR024W	SCO2	YBR037C	SCO1	0.061	0.08594	1	999	0.08452	1	0.0929	0.07787	0
YBR031W	RPL4A	YDR012W	RPL4B	NA	NA	NA	NA	NA	NA	NA	NA	NA
YBR042C	CST26	YDR018C	YDR018C	0.042	0.10686	1	0.328	0.10662	0	0.084	0.13187	1
YBR048W	RPS11B	YDR025W	RPS11A	0.083	0.05432	0	1E-04	0.05432	0	0.0492	0.05597	0
YBR049C	REB1	YDR026C	NSI1	0.034	0.07704	1	999	0.07735	1	0.0512	0.1127	1
YBR052C	RFS1	YDR032C	PST2	0.1	0.08288	0	0.605	0.07403	1	0.1019	0.04808	1
YBR054W	YRO2	YDR033W	MRH1	0.008	0.15145	1	1.077	0.15145	0	0.1596	0.13053	0
YBR066C	NRG2	YDR043C	NRG1	0.144	0.08598	1	662.6	0.08468	0	0.0692	0.10334	0
YBR068C	BAP2	YDR046C	BAP3	NA	NA	NA	NA	NA	NA	NA	NA	NA
YBR082C	UBC4	YDR059C	UBC5	0.027	0.01162	1	0.006	0.01177	0	0.0112	0.01208	0
YBR089C-A	NHP6B	YPR052C	NHP6A	0.018	0.03274	0	0.051	0.02748	0	0.0191	0.0302	0
YBR104W	YMC2	YPR058W	YMC1	0.044	0.05003	0	999	0.05079	1	0.0603	0.04529	0
YBR117C	TKL2	YPR074C	TKL1	0.155	0.06969	1	999	0.06069	1	0.1	0.03173	1
YBR118W	TEF2	YPR080W	TEF1	0.087	0.09546	0	1.127	0.09545	0	0.1042	0.0897	0
YBR140C	IRA1	YOL081W	IRA2	0.061	0.0519	1	999	0.05194	1	0.0537	0.05083	0
YBR145W	ADH5	YOL086C	ADH1	0.107	0.08104	1	999	0.0781	1	0.0661	0.25284	1
YBR147W	RTC2	YOL092W	YOL092W	0.077	0.08537	0	1.747	0.08264	0	0.1025	0.06451	1
YBR150C	TBS1	YOL089C	HAL9	0.137	0.11564	1	0.13	0.11551	0	0.0991	0.13047	1
YBR161W	CSH1	YPL057C	SUR1	0.01	0.06791	1	263.8	0.0681	0	0.0608	0.07744	0
YBR172C	SMY2	YPL105C	SYH1	0.058	0.17976	1	999	0.17811	1	0.1874	0.17043	0
YBR177C	EHT1	YPL095C	EEB1	0.047	0.08979	1	0.675	0.08752	1	0.095	0.07526	0
YBR181C	RPS6B	YPL090C	RPS6A	NA	NA	NA	NA	NA	NA	NA	NA	NA
YBR182C	SMP1	YPL089C	RLM1	0.182	0.20233	0	999	0.20099	1	0.1585	0.26812	1
YBR183W	YPC1	YPL087W	YDC1	0.04	0.06143	0	0.053	0.06155	0	0.0649	0.05797	0
YBR189W	RPS9B	YPL081W	RPS9A	0.074	0.06623	0	1E-04	0.06623	0	0.0542	0.09907	1

YBR191W	RPL21A	YPL079W	RPL21B	0.109	0.07972	0	1E-04	0.07973	0	0.0698	0.08858	0
YBR197C	YBR197C	YPL077C	YPL077C	0.07	0.13699	1	0.041	0.1375	0	0.0613	0.24815	1
YBR210W	ERV15	YGL054C	ERV14	0.031	0.0831	1	0.068	0.1403	0	0.053	0.13291	0
YBR216C	YBP1	YGL060W	YBP2	0.01	0.14286	1	0.247	0.14235	0	0.1324	0.1544	0
YBR218C	PYC2	YGL062W	PYC1	0.04	0.03775	0	999	0.0363	1	0.0388	0.0339	0
YBR238C	YBR238C	YGL107C	RMD9	0.066	0.07193	0	0.313	0.07135	0	0.1121	0.04097	1
YBR242W	YBR242W	YGL101W	YGL101W	0.028	0.06777	1	0.166	0.06718	0	0.0799	0.05434	0
YBR270C	BIT2	YJL058C	BIT61	0.009	0.07866	1	330	0.07805	0	0.0868	0.07092	0
YBR273C	UBX7	YJL048C	UBX6	0.066	0.12022	1	941.1	0.11968	0	0.0996	0.13839	1
YBR278W	DPB3	YJL065C	DLS1	0.069	0.14429	1	3.697	0.14244	0	0.1324	0.1516	0
YBR284W	YBR284W	YJL070C	YJL070C	0.005	0.10314	1	0.001	0.10371	1	0.1358	0.07259	1
YCL011C	GBP2	YNL004W	HRB1	0.062	0.08575	1	999	0.08474	1	0.0968	0.07171	1
OYCL024W	KCC4	YDR507C	GIN4	0.1	0.11534	1	0.086	0.11619	0	0.1753	0.07455	1
YCL035C	GRX1	YDR513W	GRX2	0.112	0.04965	1	0.179	0.0414	1	0.0303	0.05448	0
YCL036W	GFD2	YDR514C	YDR514C	0.014	0.11065	1	999	0.11052	0	0.1021	0.11927	0
YCL037C	SRO9	YDR515W	SLF1	0.106	0.19783	1	0.198	0.19784	0	0.1517	0.23894	1
YCL040W	GLK1	YDR516C	EMI2	0.06	0.01544	1	425.9	0.01891	1	0.0148	0.03356	1
YCL043C	PDI1	YDR518W	EUG1	0.112	0.09826	0	999	0.09643	1	0.0972	0.09563	0
YCL050C	APA1	YDR530C	APA2	0.009	0.12062	1	999	0.11697	1	0.1314	0.10217	0
CL051W	LRE1	YDR528W	HLR1	0.01	0.15459	1	530.1	0.1531	0	0.1532	0.15305	0
YCR005C	CIT2	YNR001C	CIT1	0.09	0.05138	1	999	0.04838	1	0.0436	0.05431	0
YCR010C	ADY2	YNR002C	ATO2	0.08	0.04636	1	999	0.0477	1	0.0394	0.05587	0
YCR026C	NPP1	YEL016C	NPP2	0.008	0.13528	1	0.146	0.13525	0	0.1609	0.12328	1
YCR031C	RPS14A	YJL191W	RPS14B	0.069	0.06576	0	999	0.05438	1	0.0403	0.09126	1
YCR034W	FEN1	YJL196C	ELO1	0.076	0.07142	0	0.502	0.06827	1	0.0965	0.0511	1
YCR037C	PHO87	YJL198W	PHO90	0.047	0.05944	1	999	0.059	1	0.0544	0.06338	0
YCR048W	ARE1	YNR019W	ARE2	NA	NA	NA	NA	NA	NA	NA	NA	0
YCR052W	RSC6	YNR023W	SNF12	0.071	0.13024	1	0.143	0.13014	0	0.1421	0.11723	0
YCR067C	SED4	YNR026C	SEC12	0.015	0.22398	1	999	0.22215	0	0.2923	0.09975	1
YCR069W	CPR4	YNR028W	CPR8	0.031	0.19388	1	0.106	0.19638	0	0.2602	0.14923	1
YCR073C	SSK22	YNR031C	SSK2	0.005	0.09814	1	0.045	0.10214	1	0.1392	0.02985	1
YCR073W-A	SOL2	YNR034W	SOL1	0.066	0.04246	1	0.546	0.04126	0	0.038	0.04462	0
YCR091W	KIN82	YNR047W	FPK1	0.004	0.10297	1	147.3	0.10134	1	0.1239	0.08144	1
YDL012C	YDL012C	YBR016W	YBR016W	0.083	0.08551	0	999	0.08077	1	0.1249	0.05983	1
YDL019C	OSH2	YAR042W	SWH1	0.061	0.08577	1	999	0.08405	1	0.0911	0.07752	1
YDL021W	GPM2	YOL056W	GPM3	0.122	0.07304	1	0.061	0.07442	0	0.0987	0.04961	1
YDL022W	GPD1	YOL059W	GPD2	0.088	0.07083	1	999	0.06852	1	0.0537	0.08523	1
YDL036C	PUS9	YOL066C	RIB2	0.024	0.06738	1	999	0.06676	0	0.0877	0.05398	1
YDL048C	STP4	YLR375W	STP3	0.063	0.08824	0	999	0.08759	1	0.0687	0.10164	1
YDL061C	RPS29B	YLR388W	RPS29A	0.104	0.05962	0	1E-04	0.06289	0	0.0536	0.08738	0
YDL075W	RPL31A	YLR406C	RPL31B	0.115	0.05091	1	1.176	0.05091	0	0.0461	0.0545	0
YDL082W	RPL13A	YMR142C	RPL13B	0.18	0.12353	1	1E-04	0.12423	0	0.1153	0.15524	0
YDL083C	RPS16B	YMR143W	RPS16A	0.026	0.02745	0	999	0.02569	1	0.0082	0.05508	1
YDL085W	NDE1	YMR145C	NDE1	0.082	0.06536	1	999	0.06485	1	0.0594	0.06916	0
YDL088C	ASM4	YMR153W	NUP53	0.067	0.16403	1	999	0.16253	0	0.1386	0.19151	1
YDL109C	YDL109C	YGL144C	ROG1	0.052	0.10766	1	0.279	0.10695	0	0.1571	0.06691	1
YDL137W	ARF2	YDL192W	ARF1	0.019	0.01519	0	1.271	0.01519	0	0.0143	0.01644	0
YDL161W	ENT1	YLR206W	ENT2	0.071	0.09266	1	0.157	0.09053	0	0.0914	0.08926	0
YDL179W	PCL9	YDL127W	PCL2	0.144	0.13032	0	0.015	0.13337	0	0.1225	0.14442	0
YDL181W	INH1	YDL130W-A	STF1	0.072	0.07248	0	999	0.06815	0	0.0727	0.06087	0
YDL182W	LYS20	YDL131W	LYS21	0.038	0.04438	0	999	0.04285	1	0.0418	0.044	0
YDL184C	RPL41A	YDL133C-A	RPL41B	0.011	0.27991	1	999	0.0001	1	0.0001	0.50187	0
YDL191W	RPL35A	YDL136W	RPL35B	0.108	0.13262	0	999	0.11873	1	0.0294	0.22725	1
YDL204W	RTN2	YDR233C	RTN1	0.196	0.14958	1	999	0.14721	1	0.1457	0.14891	0
YDL211C	YDL211C	YNL176C	TDA7	0.013	0.26598	1	0.053	0.26843	0	0.2428	0.29318	0
YDL214C	PRR2	YNL183C	NPR1	0.074	0.10509	1	0.376	0.10381	0	0.1738	0.04836	1
YDL222C	FMP45	YNL194C	YNL194C	0.115	0.08945	0	999	0.09029	0	0.0959	0.08278	0
YDL224C	WHI4	YNL197C	WHI3	0.116	0.10446	0	999	0.10281	1	0.1096	0.09649	0
YDL226C	GCS1	YNL204C	SPS18	0.064	0.08265	0	0.015	0.08527	1	0.1041	0.07399	1
YDL229W	SSB1	YNL209W	SSB2	0.087	0.05071	1	1E-04	0.05071	0	0.0438	0.05691	0
YDL239C	ADY3	YNL225C	CNM67	0.137	0.13427	0	999	0.13452	0	0.1544	0.11168	1
YDR028C	REG1	YBR050C	REG2	0.039	0.1203	1	999	0.11934	0	0.1769	0.10212	1
YDR066C	RTR2	YER139C	RTR1	0.04	0.06804	1	999	0.06619	0	0.1697	0.02992	1
YDR077W	SED1	YER150W	SPI1	0.271	0.17727	1	999	0.17523	0	0.2052	0.13404	1
YDR085C	AFR1	YER158C	YER158C	0.011	0.20025	1	0.163	0.20038	0	0.1655	0.24912	1
YDR093W	DNF2	YER166W	DNF1	0.004	0.04506	1	563.8	0.04558	1	0.0415	0.05045	1
YDR096W	GIS1	YER169W	RPH1	0.026	0.07653	1	999	0.0761	1	0.0915	0.06293	1
YDR098C	GRX3	YER174C	GRX4	0.053	0.07523	1	999	0.07368	1	0.086	0.06181	0
YDR099W	BMH2	YER177W	BMH1	0.052	0.03307	1	237.7	0.0326	0	0.0283	0.03801	0
YDR107C	TMN2	YLR083C	EMP70	0.057	0.08619	1	0.089	0.0861	0	0.1028	0.06604	1
YDR111C	ALT2	YLR089C	ALT1	0.067	0.06973	0	999	0.06806	1	0.0714	0.06146	0
YDR125C	ECM18	YLR099C	ICT1	0.045	0.07256	1	999	0.07183	0	0.0439	0.09629	1
YDR132C	YDR132C	YLR108C	YLR108C	0.007	0.12195	1	352.5	0.12008	0	0.1269	0.11378	0
YDR144C	MKC7	YLR120C	YPS1	0.141	0.17916	1	173.3	0.17576	0	0.1638	0.18463	0
YDR146C	SWI5	YLR131C	ACE2	0.117	0.13688	0	999	0.13336	1	0.1344	0.13223	0
YDR151C	CTH1	YLR136C	TIS11	0.009	0.11291	1	0.099	0.11304	0	0.0795	0.14378	1

YDR185C	UPS3	YLR168C	UPS2	0.004	0.05439	1	0.078	0.05425	0	0.0729	0.04059	1
YDR200C	VPS64	YLR238W	FAR10	0.013	0.11864	1	0.282	0.1176	0	0.1238	0.11389	0
YDR206W	EBS1	YLR233C	EST1	0.016	0.20463	1	41.91	0.2036	0	0.2234	0.18992	0
YDR222W	YDR222W	YLR225C	YLR225C	0.009	0.05881	1	0.135	0.05772	0	0.0603	0.05369	0
YDR247W	VHS1	YPL026C	SKS1	0.008	0.08458	1	24.01	0.08492	0	0.1154	0.05976	1
YDR251W	PAM1	YPL032C	SVL3	0.1	0.13212	1	933.2	0.13029	1	0.1099	0.1594	1
YDR252W	BTT1	YPL037C	EGD1	0.129	0.11949	0	999	0.11266	1	0.1182	0.10702	0
YDR263C	DIN7	YOR033C	EXO1	0.007	0.11895	1	0.079	0.11979	0	0.1226	0.1132	0
YDR264C	AKR1	YOR034C	AKR2	0.02	0.10143	1	0.086	0.10153	0	0.1368	0.08042	1
YDR272W	GLO2	YOR040W	GLO4	0.028	0.10728	1	0.068	0.10895	0	0.1184	0.09964	0
YDR273W	DON1	YOR042W	CUE5	0.057	0.1676	1	1.186	0.16548	0	0.2648	0.11469	1
YDR277C	MTH1	YOR047C	STD1	0.026	0.08653	1	999	0.08489	0	0.1077	0.06102	1
YDR300C	PRO1	YHR033W	YHR033W	0.028	0.04216	1	999	0.04123	1	0.0362	0.04537	0
YDR303C	RSC3	YHR056C	RSC30	0.007	0.1402	1	999	0.14067	1	0.183	0.08964	1
YDR304C	CPR5	YHR057C	CPR2	NA	NA	NA	NA	NA	NA	NA	NA	NA
YDR309C	GIC2	YHR061C	GIC1	0.029	0.13324	1	999	0.13242	0	0.1327	0.13218	0
YDR312W	SSF2	YHR066W	SSF1	0.044	0.05977	1	999	0.05674	1	0.0513	0.06505	0
YDR326C	YSP2	YHR080C	YHR080C	0.08	0.14019	1	999	0.13877	1	0.0914	0.17946	1
YDR341C	YDR341C	YHR091C	MSR1	0.082	0.06985	1	0.066	0.06994	0	0.0996	0.04577	1
YDR348C	PAL1	YHR097C	YHR097C	NA	NA	NA	NA	NA	NA	NA	NA	NA
YDR351W	SBE2	YHR103W	SBE22	0.01	0.1249	1	122.8	0.1248	1	0.1137	0.13594	0
YDR353W	TRR1	YHR106W	TRR2	0.063	0.06035	0	999	0.05889	0	0.0575	0.06	0
YDR358W	GGA1	YHR108W	GGA2	0.058	0.10627	1	0.085	0.10786	0	0.1283	0.07672	1
YDR368W	YPR1	YOR120W	GCY1	0.109	0.07586	1	999	0.07371	1	0.0787	0.06862	0
YDR379W	RG2A	YOR127W	RG1A	0.009	0.15618	1	268.5	0.15515	1	0.1452	0.16412	0
YDR385W	EFT2	YOR133W	EFT1	NA	NA	NA	NA	NA	NA	NA	NA	NA
YDR389W	SAC7	YOR134W	BAG7	0.084	0.10219	1	999	0.10066	1	0.0685	0.16011	1
YDR409W	SIZ1	YOR156C	NFI1	0.064	0.14458	1	999	0.1433	1	0.1221	0.16143	1
YDR418W	RPL12B	YEL054C	RPL12A	0.077	0.03029	1	999	0.02878	0	0.035	0.02615	0
YDR438W	THI74	YML018C	YML018C	0.009	0.11953	1	0.002	0.1211	0	0.1547	0.09758	1
YDR441C	APT2	YML022W	APT1	0.064	0.09842	1	0.038	0.10262	0	0.1025	0.10271	0
YDR447C	RPS17B	YML024W	RPS17A	0.099	0.06882	0	0.156	0.06681	0	0.0775	0.05783	0
R450W	RPS18A	YML026C	RPS18B	0.073	0.14001	1	1E-04	0.14128	0	0.1724	0.05628	1
YDR451C	YHP1	YML027W	YOX1	0.02	0.1234	1	0.058	0.12557	0	0.1622	0.09181	1
YDR453C	TSA2	YML028W	TSA1	0.096	0.06519	1	999	0.06171	1	0.0611	0.06255	0
YDR458C	HEH2	YML034W	SRC1	0.106	0.17435	1	999	0.17243	0	0.2019	0.15115	1
YDR463W	STP1	YHR006W	STP2	0.093	0.13845	1	999	0.13717	0	0.1141	0.16279	1
YDR471W	RPL27B	YHR010W	RPL27A	0.101	0.04504	1	1E-04	0.04504	0	0.0288	0.05093	0
YDR475C	JIP4	YOR019W	YOR019W	0.009	0.13689	1	136.5	0.13593	0	0.1183	0.1604	1
YDR480W	DIG2	YPL049C	DIG1	0.013	0.18521	1	0.242	0.18494	0	0.1664	0.21254	0
YDR483W	KRE2	YPL053C	KTR6	0.081	0.13583	1	0.328	0.13303	0	0.0978	0.12323	1
YDR490C	PKH1	YOL100W	PKH2	0.093	0.1363	1	0.274	0.13549	0	0.1547	0.12196	1
YDR492W	IZH1	YOL101C	IZH4	0.057	0.06164	0	0.077	0.06141	0	0.1068	0.04207	1
YDR500C	RPL37B	YLR185W	RPL37A	NA	NA	NA	NA	NA	NA	NA	NA	NA
YDR501W	PLM2	YLR183C	TOS4	NA	NA	NA	NA	NA	NA	NA	NA	NA
YDR502C	SAM2	YLR180W	SAM1	0.066	0.03644	1	999	0.03174	1	0.03	0.0327	0
YDR505C	PSP1	YLR177W	YLR177W	NA	NA	NA	NA	NA	NA	NA	NA	NA
YEL006W	YEA6	YIL006W	YIA6	0.009	0.08002	1	0.067	0.08076	0	0.0902	0.07434	0
YEL034W	HYP2	YJR047C	ANB1	0.082	0.10135	0	1E-04	0.10135	0	0.0826	0.12192	0
YEL039C	CYC7	YJR048W	CYC1	0.102	0.04475	1	999	0.0425	0	0.0413	0.04389	0
YEL041W	YEF1	YJR049C	UTR1	0.006	0.09227	1	196.5	0.09137	1	0.0477	0.10431	1
YEL047C	YEL047C	YJR051W	OSM1	0.045	0.07446	1	999	0.07173	1	0.0769	0.06522	0
YEL060C	PRB1	YOR003W	YSP3	0.071	0.12273	1	0.128	0.12251	0	0.1291	0.11777	0
YER019C-A	SBH2	YER087C-B	SBH1	0.032	0.1134	1	1E-04	0.12903	0	0.1555	0.06482	0
YER028C	MIG3	YGL209W	MIG2	NA	NA	NA	NA	NA	NA	NA	NA	NA
YER035W	EDC2	YGL222C	EDC1	0.13	0.19709	0	0.003	0.20381	1	0.1979	0.2133	0
YER037W	PHM8	YGL224C	SDT1	0.004	0.06515	1	0.102	0.0649	0	0.0805	0.05298	1
YER039C	HVG1	YGL225W	VRG4	0.074	0.0581	1	0.039	0.05909	0	0.058	0.05973	0
YER045C	ACA1	YIL036W	CST6	0.143	0.22197	1	0.24	0.22167	0	0.3016	0.17554	1
YER054C	GIP2	YIL045W	PIG2	0.174	0.12967	1	999	0.12845	0	0.1527	0.09414	1
YER056C-A	RPL34A	YIL052C	RPL34B	0.11	0.13574	0	1.789	0.12607	0	0.0999	0.1826	0
YER057C	HMF1	YIL051C	MMF1	0.103	0.08074	0	0.22	0.07615	0	0.0724	0.07903	0
YER062C	HOR2	YIL053W	RHR2	0.083	0.05032	1	145.3	0.05008	0	0.0572	0.04114	0
YER064C	VHR2	YIL056W	VHR1	0.066	0.12246	1	999	0.12166	1	0.1755	0.07786	1
YER067W	RG11	YIL057C	RG12	0.056	0.0234	1	999	0.02597	0	0.0213	0.02909	0
YER070W	RNR1	YIL066C	RNR3	0.053	0.06264	1	999	0.05662	1	0.0508	0.06197	1
YER074W	RPS24A	YIL069C	RPS24B	0.083	0.07636	0	0.915	0.07636	0	0.053	0.11361	0
YER081W	SER3	YIL074C	SER33	0.042	0.04108	0	999	0.03909	1	0.0424	0.03654	0
YER129W	SAK1	YGL179C	TOS3	0.06	0.1302	1	999	0.12916	1	0.1139	0.16413	1
YER131W	RPS26B	YGL189C	RPS26A	0.083	0.03884	1	1E-04	0.03972	0	0.0356	0.04126	0
YER132C	PMD1	YGL197W	MDS3	0.008	0.09496	1	999	0.0945	1	0.0635	0.12462	1
YFL004W	VTC2	YPL019C	VTC3	0.04	0.07198	1	0.237	0.06835	1	0.0725	0.06569	0
YFL018C	LPD1	YPL017C	IRC15	0.086	0.08105	0	0.086	0.08091	0	0.1108	0.04642	1
YFL034C-A	RPL22B	YLR061W	RPL22A	0.158	0.08525	1	0.145	0.07003	0	0.0654	0.08941	0
YFL048C	EMP47	YLR080W	EMP46	0.066	0.09698	1	999	0.097	0	0.1104	0.08753	0
YFR013W	IOC3	YOL017W	ESC8	0.008	0.14159	1	270.3	0.13959	0	0.2129	0.08613	1

YFR014C	CMK1	YOL016C	CMK2	0.034	0.06509	1	999	0.06399	1	0.0745	0.05328	1	
YFR015C	GSY1	YLR258W	GSY2	0.028	0.03607	1	0.288	0.03446	1	0.0289	0.03974	1	
YFR017C	IGD1	YOL024W	YOL024W	0.01	0.1852	1	0.213	0.18498	0	0.1079	0.35415	1	
YFR022W	ROG3	YOR018W	ROD1	0.061	0.10955	1	999	0.10909	1	0.1035	0.11589	0	
YFR023W	PES4	YHR015W	MIP6	0.005	0.09284	1	9.427	0.09239	0	0.0489	0.14506	1	
YFR024C-A	LSB3	YHR016C	YSC84	NA	NA	NA	NA	NA	NA	NA	NA	NA	
YFR031C-A	RPL2A	YIL018W	RPL2B	0.042	0.0769	1	999	0.07223	1	0.0745	0.0654	0	
YFR039C	YFR039C	YGL228W	SHE10	0.011	0.1364	1	0.004	0.1358	0	0.133	0.13875	0	
YFR040W	SAP155	YGL229C	SAP4	0.073	0.09112	1	999	0.0899	0	0.0609	0.14181	1	
YFR053C	HXK1	YGL253W	HXK2	0.105	0.11236	0	999	0.10736	1	0.0964	0.12144	1	
YGL031C	RPL24A	YGR148C	RPL24B	0.078	0.08046	0	0.072	0.08117	0	0.1246	0.02714	1	
YGL049C	TIF4632	YGR162W	TIF4631	0.123	0.13419	0	999	0.12828	1	0.1279	0.12876	0	
YGL063W	PUS2	YPL212C	PUS1	0.043	0.06268	1	999	0.06123	1	0.0516	0.06298	0	
YGL071W	AFT1	YPL202C	AFT2	0.007	0.12504	1	0.002	0.1255	0	0.1471	0.11218	1	
YGL076C	RPL7A	YPL198W	RPL7B	0.106	0.15551	1	1.585	0.14382	1	0.1306	0.15456	0	
YGL082W	YGL082W	YPL191C	YPL191C	0.007	0.12084	1	999	0.11896	1	0.1799	0.07481	1	
YGL084C	GUP1	YPL189W	GUP2	0.042	0.09897	1	0.115	0.09821	0	0.1434	0.05524	1	
	MF (ALPH A)												
	MF (ALPH A) 2	YPL187W		1	0.131	0.17572	0	999	0.17476	0	0.1988	0.14663	0
YGL096W	TOS8	YPL177C	CUP9	0.011	0.17371	1	68.75	0.17082	0	0.1859	0.15774	0	
YGL133W	ITC1	YPL216W	YPL216W	0.006	0.16214	1	0.548	0.16028	0	0.2569	0.1089	1	
YGL134W	PCL10	YPL219W	PCL8	0.008	0.0955	1	164.4	0.09565	0	0.0596	0.1498	1	
YGL135W	RPL1B	YPL220W	RPL1A	0.055	0.05148	0	0.269	0.04717	0	0.0067	0.07488	1	
YGL158W	RCK1	YLR248W	RCK2	0.008	0.12526	1	118.5	0.12475	0	0.1949	0.06249	1	
YGL162W	SUT1	YPR009W	SUT2	0.015	0.14536	1	999	0.14523	0	0.101	0.15037	0	
YGL166W	CUP2	YPR008W	HAA1	0.079	0.13888	1	999	0.13806	1	0.1172	0.23358	1	
YGR004W	PEX31	YLR324W	PEX30	0.007	0.08128	1	6.927	0.08011	0	0.0633	0.10068	1	
YGR008C	STF2	YLR327C	TMA10	0.05	0.04787	0	0.201	0.04376	0	0.0322	0.05524	0	
YGR010W	NMA2	YLR328W	NMA1	0.039	0.02934	1	0.033	0.02914	0	0.0187	0.03546	1	
YGR023W	MTL1	YLR332W	MID2	0.03	0.09754	1	0.002	0.09907	1	0.0659	0.12173	1	
YGR027C	RPS25A	YLR333C	RPS25B	0.106	0.12202	0	2.661	0.12202	0	0.1358	0.1092	0	
YGR032W	GSC2	YLR342W	FKS1	0.047	0.06122	1	999	0.0582	1	0.0613	0.05507	0	
YGR034W	RPL26B	YLR344W	RPL26A	0.141	0.10434	0	1E-04	0.10724	0	0.1022	0.11458	0	
YGR038W	ORM1	YLR350W	ORM2	0.049	0.05842	0	999	0.05773	1	0.0489	0.07079	0	
YGR041W	BUD9	YLR353W	BUD8	0.012	0.21613	1	105.6	0.21481	0	0.2472	0.18622	1	
YGR043C	NQM1	YLR354C	TAL1	0.133	0.08418	1	999	0.07668	1	0.0526	0.11477	1	
YGR049W	SCM4	YLR356W	ATG33	0.16	0.10993	1	0.004	0.10946	0	0.0787	0.1426	1	
YGR056W	RSC1	YLR357W	RSC2	0.055	0.06348	0	999	0.06331	1	0.0843	0.04258	1	
YGR070W	ROM1	YLR371W	ROM2	0.05	0.08877	1	0.167	0.08698	1	0.1401	0.04525	1	
YGR071C	ENV11	YLR373C	VID22	0.006	0.14121	1	0.134	0.14127	0	0.0937	0.1998	1	
YGR085C	RPL11B	YPR102C	RPL11A	0.084	0.0561	1	0.177	0.04448	1	0.0001	0.06698	1	
YGR092W	DBF2	YPR111W	DBF20	0.022	0.04148	1	999	0.04054	1	0.0346	0.04759	1	
YGR097W	ASK10	YPR115W	RGC1	0.009	0.12726	1	605.9	0.12655	0	0.0633	0.19945	1	
YGR108W	CLB1	YPR119W	CLB2	NA	NA	NA	NA	NA	NA	NA	NA	NA	
YGR118W	RPS23A	YPR132W	RPS23B	NA	NA	NA	NA	NA	NA	NA	NA	NA	
YGR121C	MEP1	YPR138C	MEP3	0.045	0.04105	0	910.1	0.0407	1	0.0389	0.04235	0	
YGR124W	ASN2	YPR145W	ASN1	0.07	0.05337	1	0.293	0.04409	1	0.0296	0.05219	1	
YGR131W	FHN1	YPR149W	NCE102	0.079	0.09041	0	0.082	0.09051	0	0.1073	0.08075	0	
YGR136W	LSB1	YPR154W	PIN3	0.096	0.0977	0	0.095	0.09787	0	0.1145	0.08407	0	
YGR138C	TPO2	YPR156C	TPO3	NA	NA	NA	NA	NA	NA	NA	NA	NA	
YGR141W	VPS62	YPR157W	TDA6	0.035	0.08225	1	0.124	0.08178	0	0.0877	0.07692	0	
YGR142W	BTN2	YPR158W	CUR1	0.052	0.19751	1	999	0.19726	0	0.1808	0.21032	0	
YGR143W	SKN1	YPR159W	KRE6	0.075	0.05985	1	999	0.05977	1	0.0594	0.06025	0	
	YGR169C		YJR005C-										
YGR169C-A	-A	YJR005C-A	A	0.029	0.04238	0	0.065	0.04211	0	0.035	0.05159	0	
YGR180C	RNR4	YJL026W	RNR2	0.055	0.07982	1	999	0.07991	1	0.1221	0.05755	1	
YGR188C	BUB1	YJL013C	MAD3	0.006	0.13929	1	0.002	0.1405	1	0.1606	0.13085	1	
YGR192C	TDH3	YJR009C	TDH2	0.148	0.13678	0	259.9	0.13331	1	0.1326	0.13359	0	
YGR197C	SNG1	YJR015W	YJR015W	0.026	0.1446	1	0.267	0.1443	0	0.1464	0.14061	0	
YGR209C	TRX2	YLR043C	TRX1	0.097	0.10271	0	0.151	0.09591	0	0.0215	0.12822	1	
YGR213C	RTA1	YLR046C	YLR046C	0.143	0.1257	0	0.073	0.12679	0	0.116	0.24408	1	
YGR214W	RPS0A	YLR048W	RPS0B	NA	NA	NA	NA	NA	NA	NA	NA	NA	
YGR221C	TOS2	YHR149C	SKG6	0.111	0.20797	1	0.201	0.20804	0	0.2716	0.15906	1	
YGR230W	BNS1	YHR152W	SPO12	0.017	0.23068	1	0.107	0.23222	0	0.1967	0.29514	0	
YGR238C	KEL2	YHR158C	KEL1	0.092	0.14643	1	282.3	0.14612	0	0.1966	0.10902	1	
YGR239C	PEX21	YHR160C	PEX18	0.013	0.23866	1	0.147	0.23908	0	0.2726	0.20931	0	
YGR241C	YAP1802	YHR161C	YAP1801	0.105	0.13145	1	999	0.12975	1	0.1193	0.1392	0	
YGR243W	FMP43	YHR162W	MPC2	0.064	0.04086	0	0.051	0.04051	0	0.0375	0.04353	0	
YGR254W	ENO1	YHR174W	ENO2	NA	NA	NA	NA	NA	NA	NA	NA	NA	
YGR256W	GND2	YHR183W	GND1	0.116	0.05675	1	0.033	0.06223	1	0.0882	0.02891	1	
YGR273C	YGR273C	YMR295C	YMR295C	0.073	0.09501	0	0.071	0.09568	0	0.1399	0.06342	1	
YGR279C	SCW4	YMR305C	SCW10	0.156	0.16671	0	999	0.163	1	0.1153	0.22253	1	
YGR283C	YGR283C	YMR310C	YMR310C	0.007	0.1286	1	0.006	0.13007	1	0.1672	0.09612	1	
YHL001W	RPL14B	YKL006W	RPL14A	0.118	0.12157	0	0.922	0.12157	0	0.0828	0.14419	0	

YHL003C	LAG1	YKL008C	LAC1	0.046	0.05516	0	999	0.05431	1	0.0525	0.05594	0
YHL012W	YHL012W	YKL035W	UGP1	0.039	0.07818	1	0.177	0.0726	1	0.1795	0.02833	1
YHL017W	YHL017W	YKL039W	PTM1	0.076	0.09547	1	0.051	0.09654	0	0.0803	0.11384	1
YHL033C	RPL8A	YLL045C	RPL8B	0.155	0.1408	0	1.036	0.12589	1	0.0693	0.22061	1
YHL034C	SBP1	YLL046C	RNP1	0.065	0.1006	1	0.009	0.10294	0	0.1467	0.06252	1
YHL035C	VMR1	YLL048C	YBT1	0.063	0.1207	1	4.233	0.11794	1	0.1511	0.07086	1
YHR001W	OSH7	YKR003W	OSH6	NA	NA	NA	NA	NA	NA	NA	NA	NA
YHR003C	YHR003C	YKL027W	YKL027W	0.027	0.08013	1	0.064	0.08124	0	0.069	0.09157	0
YHR021C	RPS27B	YKL156W	RPS27A	0.075	0.01236	1	1.135	0.01129	0	0.0127	0.0147	0
YHR030C	SLT2	YKL161C	KDX1	0.037	0.09447	1	0.44	0.09241	0	0.1825	0.06135	1
YHR047C	AAP1	YKL157W	APE2	0.053	0.08421	1	999	0.08117	1	0.0898	0.07162	1
YHR115C	DMA1	YNL116W	DMA2	0.061	0.0922	1	999	0.09116	1	0.0946	0.08913	0
YHR117W	TOM71	YNL121C	TOM70	0.09	0.09221	0	1.475	0.08983	1	0.1078	0.06672	1
YHR123W	EPT1	YNL130C	CPT1	0.067	0.07833	0	999	0.07699	1	0.0589	0.09976	1
YHR131C	YHR131C	YNL144C	YNL144C	0.071	0.16111	1	999	0.16041	1	0.1774	0.14635	1
YHR132W-A	IGO2	YNL157W	IGO1	0.043	0.04653	0	0.145	0.04441	0	0.0714	0.01868	1
YHR133C	NSG1	YNL156C	NSG2	0.007	0.1036	1	362.9	0.10271	0	0.1463	0.07042	1
YHR135C	YCK1	YNL154C	YCK2	NA	NA	NA	NA	NA	NA	NA	NA	NA
YHR139C	SPS100	YNL160W	YGP1	0.099	0.13113	0	999	0.13603	1	0.1818	0.10196	1
YHR146W	CRP1	YNL173C	MDG1	0.137	0.19158	1	999	0.18937	0	0.1783	0.20369	0
YHR155W	YSP1	YNL257C	SIP3	0.008	0.08915	1	11.14	0.08851	0	0.0734	0.10664	1
YHR203C	RPS4B	YJR145C	RPS4A	0.076	0.09209	0	0.321	0.08188	0	0.095	0.0472	0
YIL079C	AIR1	YDL175C	AIR2	NA	NA	NA	NA	NA	NA	NA	NA	NA
YIL095W	PRK1	YNL020C	ARK1	NA	NA	NA	NA	NA	NA	NA	NA	NA
YIL105C	SLM1	YNL047C	SLM2	0.059	0.0765	0	999	0.07398	1	0.0586	0.09268	1
YIL109C	SEC24	YNL049C	SFB2	0.061	0.07879	1	999	0.07751	1	0.0518	0.10151	1
YIL113W	SDP1	YNL053W	MSG5	0.009	0.17384	1	0.164	0.17398	0	0.1623	0.22506	0
YIL114C	POR2	YNL055C	POR1	0.071	0.08833	0	999	0.08312	1	0.0545	0.10057	1
YIL117C	PRM5	YNL058C	YNL058C	0.011	0.16079	1	0.002	0.163	0	0.1819	0.14493	0
YIL123W	SIM1	YNL066W	SUN4	0.15	0.17606	1	2.083	0.17398	0	0.1822	0.16548	0
YIL131C	FKH1	YNL068C	FKH2	0.094	0.08514	0	999	0.08445	1	0.0738	0.09028	0
YIL133C	RPL16A	YNL069C	RPL16B	0.091	0.14398	1	0.102	0.1474	0	0.1545	0.14077	0
YIL135C	VHS2	YNL074C	MLF3	0.011	0.12155	1	999	0.12154	0	0.1287	0.11401	0
YIL138C	TPM2	YNL079C	TPM1	0.033	0.05483	1	999	0.05322	0	0.0505	0.05613	0
YIL148W	RPL40A	YKR094C	RPL40B	0.007	0.06106	1	2.98	0.03686	1	0.1223	0.02282	1
YIL151C	YIL151C	YKR096W	YKR096W	0.008	0.09411	1	0.435	0.09343	0	0.0836	0.10089	1
YIL158W	AIM20	YKR100C	SKG1	0.047	0.1256	1	0.076	0.12617	0	0.1069	0.17829	1
YIR018W	YAP5	YOL028C	YAP7	0.012	0.13209	1	999	0.13152	0	0.1324	0.01304	0
YIR021W	MRS1	YKL011C	CCE1	0.015	0.16023	1	0.036	0.16165	0	0.2014	0.13317	1
YIR033W	MGA2	YKL020C	SPT23	0.007	0.08927	1	87.4	0.08918	1	0.11	0.07024	1
YIR037W	HYR1	YKL026C	GPX1	0.082	0.06164	0	999	0.06011	0	0.0412	0.08747	1
YJL045W	YJL045W	YKL148C	SDH1	0.061	0.04144	1	0.095	0.03857	1	0.0502	0.02258	1
YJL076W	NET1	YKR010C	TOF2	0.076	0.17934	1	999	0.17802	1	0.149	0.23624	1
YJL082W	IML2	YKR018C	YKR018C	0.006	0.04583	1	18.1	0.04517	1	0.0465	0.04375	0
YJL083W	TAX4	YKR019C	IRS4	0.078	0.1317	1	999	0.13096	1	0.1313	0.13064	0
YJL084C	ALY2	YKR021W	ALY1	0.022	0.09412	1	999	0.09392	1	0.0951	0.09251	0
YJL099W	CHS6	YKR027W	BCH2	0.053	0.11132	1	81.49	0.11116	0	0.1001	0.12189	0
YJL105W	SET4	YKR029C	SET3	0.009	0.16297	1	0.18	0.16283	0	0.1468	0.1754	0
YJL110C	GZF3	YKR034W	DAL80	0.156	0.13767	0	999	0.13699	1	0.1675	0.07485	1
YJL112W	MDV1	YKR036C	CAF4	0.018	0.14028	1	2.317	0.1393	0	0.1598	0.12247	1
YJL129C	TRK1	YKR050W	TRK2	0.006	0.12943	1	42.4	0.12735	0	0.1406	0.11783	1
YJL133W	MRS3	YKR052C	MRS4	0.027	0.03564	0	999	0.03801	0	0.0283	0.04608	0
YJL134W	LCB3	YKR053C	YSR3	0.068	0.09426	1	999	0.09318	0	0.1185	0.07416	1
YJL137C	GLG2	YKR058W	GLG1	0.137	0.21314	1	0.085	0.22158	1	0.159	0.25972	1
YJL138C	TIF2	YKR059W	TIF1	0.042	0.04874	0	0.145	0.0487	0	0.0365	0.15513	1
YJL159W	HSP150	YKL163W	PIR3	0.163	0.12681	1	39.67	0.1158	1	0.093	0.13358	1
YJL164C	TPK1	YKL166C	TPK3	0.041	0.05293	1	999	0.05254	1	0.0485	0.05687	0
YJL165C	HAL5	YKL168C	KKQ8	0.008	0.09504	1	999	0.09506	1	0.1338	0.06408	1
YJL177W	RPL17B	YKL180W	RPL17A	0.11	0.04785	1	1E-04	0.04842	0	0.0381	0.06499	0
YJL181W	YJL181W	YJR030C	YJR030C	0.011	0.22981	1	0.009	0.23407	1	0.3315	0.18426	1
YJR054W	YJR054W	YML047C	PRM6	0.008	0.11109	1	0.024	0.11228	0	0.1051	0.11682	0
YJR059W	PTK2	YKL198C	PTK1	0.067	0.09251	0	93.15	0.09089	0	0.0825	0.10064	0
YJR066W	TOR1	YKL203C	TOR2	0.018	0.02981	1	358.3	0.03008	1	0.0352	0.02585	1
YJR091C	JSN1	YPR042C	PUF2	NA	NA	NA	NA	NA	NA	NA	NA	NA
YJR094W-A	RPL43B	YPR043W	RPL43A	NA	NA	NA	NA	NA	NA	NA	NA	NA
YJR127C	RSF2	YML081W	TDA9	0.119	0.0964	1	999	0.09604	1	0.1088	0.08231	1
YJR130C	STR2	YML082W	YML082W	0.062	0.04809	1	999	0.0482	1	0.0453	0.05103	0
YJR148W	BAT2	YHR208W	BAT1	0.086	0.06821	1	999	0.06118	1	0.0504	0.06968	1
YKL032C	IXR1	YMR072W	ABF2	0.129	0.11388	0	999	0.10931	1	0.1237	0.10429	0
YKL043W	PHD1	YMR016C	SOK2	0.272	0.16904	1	999	0.16854	0	0.161	0.17199	0
YKL050C	YKL050C	YMR031C	EIS1	0.069	0.16575	1	0.751	0.1638	0	0.2083	0.12576	1
YKL065C	YET1	YMR040W	YET2	0.067	0.10632	1	0.005	0.11179	1	0.0054	0.11867	1
YKL072W	STB6	YMR053C	STB2	0.008	0.07352	1	9.404	0.07356	0	0.0788	0.06853	0
YKL091C	YKL091C	YMR079W	SEC14	0.041	0.04049	0	999	0.0375	1	0.0361	0.0392	0
YKL093W	MBR1	YMR081C	ISF1	0.01	0.16353	1	999	0.1623	0	0.1603	0.16421	0
YKL105C	YKL105C	YMR086W	SEG1	0.009	0.249	1	0.058	0.25047	0	0.2646	0.23514	0

YKL121W	DGR2	YMR102C	YMR102C	0.007	0.08378	1	18.85	0.0835	1	0.1105	0.05849	1
YKL127W	PGM1	YMR105C	PGM2	0.035	0.06218	1	999	0.05768	1	0.0501	0.06423	1
YKL129C	MYO3	YMR109W	MYO5	0.047	0.05736	1	999	0.05579	1	0.0677	0.04039	1
YKL132C	RMA1	YMR113W	FOL3	0.005	0.12508	1	0.105	0.12585	0	0.1703	0.09161	1
YKL133C	YKL133C	YMR115W	MGR3	NA	NA	NA	NA	NA	NA	NA	NA	NA
YKL141W	SDH3	YMR118C	SHH3	0.162	0.10184	1	0.17	0.10065	0	0.1658	0.06714	1
YKL187C	YKL187C	YLR413W	YLR413W	0.124	0.14617	0	999	0.1431	1	0.1726	0.1208	1
YKL201C	MNN4	YJR061W	YJR061W	0.006	0.09528	1	1.959	0.09459	0	0.0832	0.10959	1
YKR072C	SIS2	YOR054C	VHS3	0.046	0.10342	1	999	0.10232	0	0.0955	0.11166	0
YKR075C	YKR075C	YOR062C	YOR062C	0.053	0.13173	1	0.255	0.13135	0	0.167	0.10557	1
YKR077W	MSA2	YOR066W	MSA1	0.009	0.13819	1	999	0.13806	0	0.1365	0.13892	0
YKR078W	YKR078W	YOR069W	VPS5	NA	NA	NA	NA	NA	NA	NA	NA	NA
YKR084C	HBS1	YOR076C	SKI7	NA	NA	NA	NA	NA	NA	NA	NA	NA
YKR089C	TGL4	YOR081C	TGL5	0.009	0.05608	1	19.48	0.05704	1	0.0612	0.05608	0
YKR091W	SRL3	YOR083W	WHI5	0.024	0.17256	1	0.147	0.1727	0	0.2062	0.14459	1
YLL010C	PSR1	YLR019W	PSR2	0.078	0.14401	1	999	0.14306	0	0.177	0.10744	1
YLL012W	YEH1	YLR020C	YEH2	0.005	0.06292	1	28.83	0.0635	1	0.0593	0.06794	0
YLL021W	SPA2	YLR313C	SPH1	0.102	0.2197	1	101.5	0.21546	1	0.1927	0.29761	1
YLR028C	ADE16	YMR120C	ADE17	NA	NA	NA	NA	NA	NA	NA	NA	NA
YLR031W	YLR031W	YMR124W	YMR124W	0.123	0.15092	1	0.626	0.14714	1	0.1648	0.14618	0
YLR174W	IDP2	YNL009W	IDP3	0.037	0.04118	0	256.7	0.03997	0	0.0517	0.02539	1
YLR187W	SKG3	YNL278W	CAF120	0.122	0.15257	1	917.6	0.14976	1	0.1338	0.17152	1
YLR249W	YEF3	YNL014W	HEF3	0.072	0.06405	1	0.548	0.06109	1	0.0527	0.09252	1
YLR260W	LCB5	YOR171C	LCB4	0.006	0.08556	1	43.98	0.08508	1	0.0842	0.08601	0
YLR264W	RPS28B	YOR167C	RPS28A	0.014	0.02299	0	1E-04	0.02321	0	0.0284	0.01629	0
YLR270W	DCS1	YOR173W	DCS2	NA	NA	NA	NA	NA	NA	NA	NA	NA
YLR277C	YSH1	YOR179C	SYC1	0.021	0.0672	1	0.042	0.06764	0	0.1794	0.05409	1
YLR287C-A	RPS30A	YOR182C	RPS30B	0.076	0.14443	1	1E-04	0.14915	0	0.1645	0.13623	0
YLR293C	GSP1	YOR185C	GSP2	0.011	0.01093	0	0.02	0.00935	0	0.0051	0.01376	0
YLR300W	EXG1	YOR190W	SPR1	0.085	0.10678	1	999	0.10451	1	0.1041	0.105	0
YLR441C	RPS1A	YML063W	RPS1B	0.07	0.06008	0	1E-04	0.06007	0	0.0615	0.0573	0
YLR442C	SIR3	YML065W	ORC1	0.005	0.13357	1	0.076	0.13406	0	0.2086	0.07038	1
YLR448W	RPL6B	YML073C	RPL6A	0.104	0.14001	0	5.278	0.13842	0	0.1071	0.18185	1
YLR449W	FPR4	YML074C	FPR3	0.063	0.1201	1	999	0.11811	1	0.1306	0.10458	0
YLR450W	HMG2	YML075C	HMG1	0.047	0.05713	1	0.157	0.05536	1	0.044	0.06541	1
YLR456W	YLR456W	YPR172W	YPR172W	0.008	0.07602	1	7.035	0.07479	0	0.1073	0.04276	1
YLR457C	NBP1	YPR174C	YPR174C	0.01	0.11959	1	2E-04	0.12209	0	0.1115	0.12898	0
YML007W	YAP1	YDR423C	CAD1	0.092	0.17259	1	0.296	0.17199	0	0.1224	0.27082	1
YML056C	IMD4	YLR432W	IMD3	0.065	0.0863	1	0.244	0.07873	1	0.0369	0.10471	1
YML106W	URA5	YMR271C	URA10	0.035	0.04255	0	999	0.03887	0	0.0733	0.02382	1
YML109W	ZDS2	YMR273C	ZDS1	0.069	0.17016	1	0.227	0.16958	0	0.181	0.15993	0
YML111W	BUL2	YMR275C	BUL1	0.012	0.06845	1	246.8	0.06843	1	0.068	0.06869	0
YML116W	ATR1	YMR279C	YMR279C	0.077	0.08939	1	999	0.08759	1	0.0777	0.09756	1
YML118W	NGL3	YMR285C	NGL2	0.033	0.09544	1	0.123	0.09503	0	0.1383	0.06056	1
YML124C	TUB3	YML085C	TUB1	0.022	0.02311	0	5.504	0.02311	0	0.0228	0.02371	0
YML125C	PGA3	YML087C	AIM33	0.067	0.07368	0	0.174	0.07203	0	0.0599	0.08143	0
YMR008C	PLB1	YOL011W	PLB3	0.081	0.06984	1	0.256	0.06831	1	0.0658	0.07067	0
YMR119W	ASI1	YNL008C	ASI3	0.009	0.16746	1	0.088	0.16811	0	0.143	0.19639	1
YMR175W	SIP18	YPL223C	GRE1	0.278	0.09333	1	0.013	0.0945	0	0.0911	0.09621	0
YMR177W	MMT1	YPL224C	MMT2	0.058	0.12159	1	0.558	0.11982	0	0.1075	0.12993	0
YMR180C	CTL1	YPL228W	CET1	0.034	0.1307	1	0.038	0.13212	0	0.2609	0.09048	1
YMR181C	YMR181C	YPL229W	YPL229W	0.013	0.11213	1	999	0.11197	0	0.1188	0.10063	0
YMR182C	RGM1	YPL230W	USV1	0.011	0.20655	1	1.039	0.20504	0	0.2343	0.1499	1
YMR183C	SSO2	YPL232W	SSO1	0.056	0.04537	0	999	0.04492	1	0.0406	0.04886	0
YMR186W	HSC82	YPL240C	HSP82	0.09	0.04844	1	0.184	0.04556	1	0.0549	0.02486	1
YMR194W	RPL36A	YPL249C-A	RPL36B	0.089	0.10326	0	0.041	0.13404	0	0.0776	0.22986	0
YMR198W	CIK1	YPL253C	VIK1	0.012	0.12245	1	0.002	0.12427	1	0.1137	0.13548	0
YMR199W	CLN1	YPL256C	CLN2	0.01	0.06391	1	82.38	0.06438	1	0.0664	0.06197	0
YMR206W	YMR206W	YNR014W	YNR014W	0.01	0.14771	1	0.063	0.14882	0	0.1528	0.14688	0
YMR207C	HFA1	YNR016C	ACC1	0.061	0.07718	1	0.146	0.07268	1	0.0216	0.09776	1
YMR230W	RPS10B	YOR293W	RPS10A	NA	NA	NA	NA	NA	NA	NA	NA	NA
YMR242C	RPL20A	YOR312C	RPL20B	0.11	0.10789	0	1E-04	0.1122	0	0.1172	0.10935	0
YMR243C	ZRC1	YOR316C	COT1	0.063	0.09424	1	999	0.09235	0	0.0724	0.11309	1
YMR246W	FAA4	YOR317W	FAA1	NA	NA	NA	NA	NA	NA	NA	NA	NA
YMR264W	CUE1	YML101C	CUE4	0.079	0.09867	0	999	0.09577	0	0.2103	0.07418	1
YNL001W	DOM34	YCL001W-B	YCL001W-B	0.013	0.04692	1	489	0.04567	0	0.0136	0.04818	1
YNL052W	COX5A	YIL111W	COX5B	0.159	0.08013	1	0.156	0.0631	1	0.0577	0.06752	0
YNL065W	AQR1	YIL120W	QDR1	0.034	0.07209	1	1.94	0.07033	1	0.0925	0.04877	1
YNL087W	TCB2	YOR086C	TCB1	NA	NA	NA	NA	NA	NA	NA	NA	NA
YNL093W	YPT53	YOR089C	VPS21	0.028	0.0809	1	0.02	0.08431	1	0.1442	0.04658	1
YNL095C	YNL095C	YOR092W	ECM3	0.085	0.09342	0	999	0.09308	1	0.0766	0.10673	1
YNL096C	RPS7B	YOR096W	RPS7A	0.125	0.10643	0	1E-04	0.10978	0	0.0887	0.15955	1
YNL098C	RAS2	YOR101W	RAS1	0.044	0.11013	1	9.264	0.11096	1	0.0877	0.15519	1
YNL104C	LEU4	YOR108W	LEU9	0.051	0.05734	0	999	0.05425	1	0.0478	0.0594	0
YNL106C	INP52	YOR109W	INP53	0.039	0.07849	1	699	0.07627	1	0.0481	0.10108	1

YNL108C	YNL108C	YOR110W	TFC7	0.015	0.08832	1	0.101	0.08818	0	0.1049	0.06093	1
YNL293W	MSB3	YOL112W	MSB4	0.01	0.0796	1	1.035	0.07903	0	0.0811	0.07769	0
YNL298W	CLA4	YOL113W	SKM1	0.087	0.07703	0	337.2	0.07561	1	0.0864	0.06871	0
YNL299W	TRF5	YOL115W	PAP2	0.007	0.08146	1	550.2	0.08027	1	0.0762	0.08418	0
YNL301C	RPL18B	YOL120C	RPL18A	NA	NA	NA	NA	NA	NA	NA	NA	NA
YNL302C	RPS19B	YOL121C	RPS19A	0.097	0.11607	0	1E-04	0.12157	0	0.1441	0.09296	0
YNL309W	STB1	YOL131W	YOL131W	0.009	0.13333	1	0.331	0.13262	0	0.1972	0.12028	1
YNR034W-A	YNR034W	YCR075W-A	YCR075W-	0.003	0.06647	1	0.006	0.06882	0	0.0504	0.09373	0
YOL036W	-A	YIR016W	A	0.009	0.20292	1	0.153	0.203	0	0.183	0.29917	1
YOL091W	YOL036W	YIR016W	YIR016W	0.012	0.15105	1	999	0.1511	0	0.2076	0.10629	1
YOR016C	SPO21	YBR148W	YSW1	0.072	0.0857	0	0.063	0.08591	0	0.0556	0.10728	1
YOR044W	ERP4	YAL007C	ERP2	0.013	0.16269	0	0.271	0.16129	0	NA	NA	NA
YOR222W	IRC23	YDR275W	BSC2	0.048	0.05323	0	999	0.0557	1	0.0625	0.05087	0
YOR226C	ODC2	YPL134C	ODC1	0.071	0.04216	1	999	0.04426	1	0.0506	0.03954	0
YOR227W	ISU2	YPL135W	ISU1	0.101	0.11852	1	999	0.11657	1	0.1153	0.11808	0
YOR229W	HER1	YPL137C	GIP3	0.025	0.09129	1	0.095	0.09124	0	0.0427	0.16137	1
YOR231W	WTM2	YPL139C	UME1	0.004	0.08913	1	12.24	0.08842	0	0.0879	0.08899	0
YOR233W	MKK1	YPL140C	MKK2	0.089	0.1482	1	0.364	0.14592	0	0.1203	0.17828	1
YOR247W	KIN4	YPL141C	FRK1	0.047	0.2276	0	0.001	0.23828	1	NA	NA	NA
YOR256C	SRL1	YPL163C	SVS1	0.047	0.11531	1	999	0.1143	0	0.1321	0.0995	1
YPL112C	TRE2	YPL176C	TRE1	0.007	0.13701	1	0.001	0.14073	1	0.2766	0.09878	1
YPL250C	PEX25	YOR193W	PEX27	0.03	0.20508	1	0.036	0.21446	0	0.2453	0.18844	0
	ICY2	YMR195W	ICY1									

ANEXO C

Oligonucleótidos de clonación en el vector pFA6

Nombre	Secuencia
ERV14/XmaI_X1fo	CCGGAcccgggattccaacaatctcaacattcacc
ERV14/XmaI_X1re	CCGGAcccgggggaagcacgaattaatacattctgtg
ERV14/XmaI_X2fo	CCGGAcccgggatcataaccgaaggatgaccaactt
ERV14/XmaI_X2re	CCGGAcccgggtggatccatataacggacttctgaa
TSA1/XmaI_X1fo	CCGGAcccgggtgctatcgggtgagattcctaagg
TSA1/XmaI_X1re	CCGGAcccgggcggtagtgtatcgttttgtcagtga
TSA1/XmaI_X2fo	CCGGAcccgggatattgatatgcagggtgtcaaactgtc
TSA1/XmaI_X2re	CCGGAcccgggggtcatttgctctctctggatataaagatg
RHR2/XmaI_X1fo	CCGGAcccggggacgtatggctctgggtgcaagatc
RHR2/XmaI_X1re	CCGGAcccggggcgaagacgacaaagaaaaggatgc
RHR2/XmaI_X2fo	CCGGAcccggggagcaatgtctacgtgaaagataccaa
RHR2/XmaI_X2re	CCGGAcccggggggaatataagaataggactgtatctgaga
FEN1/XmaI_X1fo	CCGGAcccggggaggcactcctactttccctttaccg
FEN1/XmaI_X1re	CCGGAcccggggactgtcgaaagaaacggaaaact
FEN1/XmaI_X2fo	CCGGAcccggggctggcaatgcaccatcgctgtatt
FEN1/XmaI_X2re	CCGGAcccgggtatgtttttgtcctgggcgcgacatt
GUP1/XmaI_X1fo	CCGGAcccggggcactgttttgaataactttgtgctc
GUP1/XmaI_X1re	CCGGAcccggggggactgtggatccagaataaatttga
GUP1/XmaI_X2fo	CCGGAcccgggggttaaattttcctaattgtgtctcgaggtgc
GUP1/XmaI_X2re	CCGGAcccggggcagcattttctgacagttacagttttc
SAC7/XmaI_X1fo	CCGGAcccgggatataccaaaagggaaggagcgcgt
SAC7/XmaI_X1re	CCGGAcccggggcgcaccagtaacaaaaggacaaagt
SAC7/XmaI_X2fo	CCGGAcccgggacgtatcaccacccggttagatatac
SAC7/XmaI_X2re	CCGGAcccggggaataggactgaccaaccagagcgg
BUL2/XmaI_X1fo	CCGGAcccggggttagtataaaacagtcaatttctgtggacctgg
BUL2/XmaI_X1re	CCGGAcccggggcagaattcgtatatactaggggaaatagaacga
BUL2/XmaI_X2fo	CCGGAcccggggcacaacatgtctgcagatggcatat
BUL2/XmaI_X2re	CCGGAcccggggcaacaaaagagcaccagaaaatgcat
ITR1/XmaI_X1fo	CCGGAcccgggatgtctatttatcacttgggttctctcc
ITR1/XmaI_X1re	CCGGAcccgggccttcgaaatacttttaggtaggcatctc
ITR1/XmaI_X2fo	CCGGAcccgggggcttacttactgatagtagatcaacgatc
ITR1/XmaI_X2re	CCGGAcccggggaaagaacataaccgctcgaaaaga
TIF4631/XmaI_X1fo	CCGGAcccggggctttttctctgttctgtttcaag
TIF4631/XmaI_X1re	CCGGAcccggggcacagtaagaggattaccatataaaatgc
TIF4631/XmaI_X2fo	CCGGAcccgggaaaggtgtctgtacagaagctggatg
TIF4631/XmaI_X2re	CCGGAcccgggggacataccaagaattactcagccttt
SET3/XmaI_X1fo	CCGGAcccggggaatcaaatggaacatttcccagctg
SET3/XmaI_X1re	CCGGAcccggggctccactttttctgcatataggtcat
SET3/XmaI_X2fo	CCGGAcccggggcagcaaggcacactcgtatagcatgt
SET3/XmaI_X2re	CCGGAcccgggggtggattttaacgtaaggatcctctcaact
CLA4/XmaI_X1fo	CCGGAcccggggccacatcgatgtccaagtgaacaa
CLA4/XmaI_X1re	CCGGAcccgggaaaattcgacccttggaacaaaca
CLA4/XmaI_X2fo	CCGGAcccgggctatttaccttattaacgttttggcctcctgg
CLA4/XmaI_X2re	CCGGAcccggggcgaagcaaacgtatcattgttgttt
VPS5/XmaI_X1fo	CCGGAcccggggcattgttggcaaacgaactttcc
VPS5/XmaI_X1re	CCGGAcccggggatgtacgtcgcactggctgtaaa
VPS5/XmaI_X2fo	CCGGAcccgggttcacgtctcctaacaacaaccgctg
VPS5/XmaI_X2re	CCGGAcccggggcagagatatacccaattgcccacat
SDH3/XmaI_X1fo	CCGGAcccgggTTGAAAACCTTCTGGAAGGTAAGGTGGA
SDH3/XmaI_X1re	CCGGAcccgggCGCCTATGTTTGCTTTTCTTGCCAGTA

SDH3/XmaI_X2fo	CCGGAcccgggGGCTATAATGTTGTGTATAC
SDH3/XmaI_X2re	CCGGAcccgggTTATCCTTTTGTGACATGGCTCTTGCG
GSP1/XmaI_X1fo	CCGGAcccgggTGCTGAGGAAACCAAGAAAGTCATTGA
GSP1/XmaI_X1re	CCGGAcccgggGCTCCTACATCCGTTTGAAACGATCCA
GSP1/XmaI_X2fo	CCGGAcccgggAACTTACGATCCCTCTGCTTGCGGA
GSP1/XmaI_X2re	CCGGAcccgggCGGATCCCGTTTATTGGTCTTTGATTC
PET9/XmaI_X1fo	CCGGAcccgggACAAAAGCAGCCCAACTGATAACGAAAT
PET9/XmaI_X1re	CCGGAcccgggCTGCAAAATCCAAGGAACACACAAATGT
PET9/XmaI_X2fo	CCGGAcccgggCGATGCCTATCGTAGTTTGCAAGGAAGA
PET9/XmaI_X2re	CCGGAcccgggCAACCATCTGAAAGCCGAAAGTTAAAGG
RSC6/XmaI_X1fo	CCGGAcccgggGTTTCGTCGAGAACGGTGAAGATGGTG
RSC6/XmaI_X1re	CCGGAcccgggGGGATGGTGGCCGTC TAAGAACAAA
RSC6/XmaI_X2fo	CCGGAcccgggAAGACCGGGGATCTTAACCTCTTCGA
RSC6/XmaI_X2re	CCGGAcccgggCATTTGTTGCTTTTCCCTTGTTCCCTT
CDC19/XmaI_X1fo	CCGGAcccgggAACGGCGGGATTCCTCTATGGCGT
CDC19/XmaI_X1re	CCGGAcccgggGAAGATTATCTTGCGATGGGAGGGA
CDC19/XmaI_X2fo	CCGGAcccgggGGACGCCTTGTGTAACGCTTTTATGA
CDC19/XmaI_X2re	CCGGAcccgggCCAGACTGTGCGTAAACTTGACATGGA
PDI1/XmaI_X1fo	CCGGAcccgggTGGCTTGTGTAAGTCGTCGAATGACA
PDI1/XmaI_X1re	CCGGAcccgggGGAGAGGGAATGGAGGAGGATGAGATAA
PDI1/XmaI_X2fo	CCGGAcccgggTCACAGTTCATATAGGTCCCACGGC
PDI1/XmaI_X2re	CCGGAcccgggGGGAACCTGCAAGAATGGTGGAAAA
DED1/XmaI_X1fo	CCGGAcccgggAGGATTTGCGCCTTTGGATGAGG
DED1/XmaI_X1re	CCGGAcccgggGCCAATCAACGGCTAGAGAGAAACAAAC
DED1/XmaI_X2fo	CCGGAcccgggCCGCCTTAGGTTCGAATGTGGTAATF
DED1/XmaI_X2re	CCGGAcccgggCCCATTTATGATTGGGACAAGAGCAAAA
REB1/XmaI_X1fo	CCGGAcccgggCGCTGAAACAGAAAATGACATTGACGC
REB1/XmaI_X1re	CCGGAcccgggCCAGTCCTTCCAAGAATGTAGGCAGAAG
REB1/XmaI_X2fo	CCGGAcccgggGCATCTGTACACAACCTGGAAACTCA
REB1/XmaI_X2re	CCGGAcccgggGCTGCTGCTGGTAAGGCTAACAAACAA

ANEXO D

Oligonucleótidos de integración en genoma

Nombre	Secuencia
VPS5/X1pr.re	tagattatcctcgtagtcocattggtt
VPS5/X1orf.fo	taacaagccatggactacgaggataatcta
VPS5/X1 (-600f)_R3	tttgagcagggattttataaaactttcatacatcctgcCATCGATGAATTCGAGCTCGTT
VPS5/X1 (-45f)_X2orf.fo	gggattttataaaactttcatacatcctgcaataacaagccATGGAGAACGACAAGGCGTCACATGCGAGC
VPS5/X1 (-45f)_F1	gggattttataaaactttcatacatcctgcaataacaagccCCAGCTGAAGCTTCGTACGC
VPS5/X1 (1r)_X2pr.re	tagttcgtcccaaacaggtgcttctagattatcctcgtagtcocatATCTAATCTGTAGGGTTCGAGTTCTGATCT
VPS5/X1 (1r)_F1	cgccccaaacaggtgcttctagattatcctcgtagtcocatCCAGCTGAAGCTTCGTACGC
VPS5/X1 (+50r)_R3	ggacctattcattttacttactgtatagatgtaataaCATCGATGAATTCGAGCTCGTT
TSA1/X1pr.re	tgaacttgagcgaccattgtgtatgtatg
TSA1/X1orf.fo	acatacatacacaatggctcgtcaagtcca
TSA1/X1 (-600f)_R3	attgctgctgatgatttggaggagtcocaaagaagCATCGATGAATTCGAGCTCGTT
TSA1/X1 (-45f)_X2orf.fo	tgctcacaaccaaccacaactacatacatacatacacaATGGTAGCAGAAGTCAAAAACAGCC
TSA1/X1 (-45f)_F1	tgctcacaaccaaccacaactacatacatacatacacaCCAGCTGAAGCTTCGTACGC
TSA1/X1 (1r)_X2pr.re	agttttcttaaaagtggagcttggcttttgaacttgagcgaccatATTGGTTTTTACGTTCTTGTAAAGGCTAAT
TSA1/X1 (1r)_F1	tcttaaaagtggagcttggcttttgaacttgagcgaccatCCAGCTGAAGCTTCGTACGC
TSA1/X1 (+50r)_R3	cattaagtttatgggttttatatatagatagataggtacCATCGATGAATTCGAGCTCGTT
TIF4631/X1pr.re	cagtttcgtctgtcattattgttaataggt
TIF4631/X1orf.fo	tacctattacaataatgacagacgaaactg
TIF4631/X1 (-600f)_R3	gtttcatgtgacctattccccttaattggatcaatctCATCGATGAATTCGAGCTCGTT
TIF4631/X1 (-45f)_X2orf.fo	aagcaaggttaagaggacaactgtaattacctattacaataATGACTGACCAAAGAGGTCCACCGCCCCA
TIF4631/X1 (-45f)_F1	aagcaaggttaagaggacaactgtaattacctattacaataCCAGCTGAAGCTTCGTACGC
TIF4631/X1 (1r)_X2pr.re	ctgctttgaagcagattgtgtcgggtgagcagtttcgtctgtcatAGTGCCTACAATTGATCTATTGTTCTTTGT
TIF4631/X1 (1r)_F1	ttgaagcagattgtgtcgggtgagcagtttcgtctgtcatCCAGCTGAAGCTTCGTACGC
TIF4631/X1 (+50r)_R3	tagcttttagtgaaagaatgagccattaatcgagaCATCGATGAATTCGAGCTCGTT
SET3/X1pr.re	tttggaaataggtactgacatgattcagtt
SET3/X1orf.fo	aactgaatcatgtcagtaacctaatccaaa
SET3/X1 (-600f)_R3	tatacagaaagttagtaataaaggacctttctagtaaacCATCGATGAATTCGAGCTCGTT
SET3/X1 (-45f)_X2orf.fo	tacaacagtttttagatcgtacttcacaaaatacagaactgaatcATGACTTACCCGGAATCACTATCTTCTCGT
SET3/X1 (-45f)_F1	cagtttttagatcgtacttcacaaaatacagaactgaatcCCAGCTGAAGCTTCGTACGC
SET3/X1 (1r)_X2pr.re	tgcatcatctaggagcactgctcctttggaataggtactgacatTTTGAAGATGAAGCGGTATTGTAACTTTC
SET3/X1 (1r)_F1	catctaggagcactgctcctttggaataggtactgacatCCAGCTGAAGCTTCGTACGC
SET3/X1 (+50r)_R3	atgaatattcacttttgaataacttaagtttatatagCATCGATGAATTCGAGCTCGTT
SDH3/X1pr.re	ctaaactgaccatcattgcagacatcgtt
SDH3/X1orf.fo	aacgatgtctgcaatgatggtcaagttag
SDH3/X1 (-600f)_R3	caggttctagaaatccctcacttatatatatatatctatCATCGATGAATTCGAGCTCGTT
SDH3/X1 (-45f)_X2orf.fo	gaccaacccaaaaatgccaatcacaagctcctaagaataataaacgATGAAAGCAACCATTCAAAGAGTAACATCTG
SDH3/X1 (-45f)_F1	gaccaacccaaaaatgccaatcacaagctcctaagaataataaacgCCAGCTGAAGCTTCGTACGC
SDH3/X1 (1r)_X2pr.re	tagagccgactttatcagtcctaacttgaccatcattgcagacatTATTCTTACGCTTCTGAATATTCTGAGCTC
SDH3/X1 (1r)_F1	tagagccgactttatcagtcctaacttgaccatcattgcagacatCCAGCTGAAGCTTCGTACGC
SDH3/X1 (+50r)_R3	aaataagtgcggattgggtcatataggaacgtatataaaaaCATCGATGAATTCGAGCTCGTT
SAC7/X1pr.re	gtttaagagtattatttggcattataaaga
SAC7/X1orf.fo	tctttataatgccaataataactcttaaac
SAC7/X1 (-600f)_R3	ctaaaacgacgttctcagtcctctctcttttataaCATCGATGAATTCGAGCTCGTT
SAC7/X1 (-45f)_X2orf.fo	ttcatcctcacaataaccactaagataaaactaaaaatattctttataATGTTTAAACATGAATCTACTTTCTACACCT
SAC7/X1 (-45f)_F1	cctctcaataaccactaagataaaactaaaaatattctttataCCAGCTGAAGCTTCGTACGC
SAC7/X1 (1r)_X2pr.re	aacattttcaattttggagccttggtaagagtattatttggcatTCTATTTTATGTTGTGTAATAAGATATGTC
SAC7/X1 (1r)_F1	tttcaattttggagccttggtaagagtattatttggcatCCAGCTGAAGCTTCGTACGC

SAC7/X1 (+50r)_R3 ccctttatggttataaatcttgaagggaatgatgttCATCGATGAATTCGAGCTCGTT
RSC6/X1pr.re gattggtctgtgttaccatttctatgatgctg
RSC6/X1orf.fo cagctcatagaaatggttaacacagaccaatc
RSC6/X1 (-600f)_R3 gctacatgaactactcttcttctctacatagttcttcccttatgCATCGATGAATTCGAGCTCGTT
RSC6/X1 (-45f)_X2orf.fo aggcagcttctgtatagtaataacacagcagaaaaacagctcatagaaATGTCCAAGTAATGAAACCAGCAACG
RSC6/X1 (-45f)_F1 aggcagcttctgtatagtaataacacagcagaaaaacagctcatagaaCCAGCTGAAGCTTCGTACGC
RSC6/X1 (1r)_X2pr.re atccggttgatagtaaacagggaccggattggtctgtgttaccatGGTTCAATCCGTTAGATCTCTGTTTG
RSC6/X1 (1r)_F1 atccggttgatagtaaacagggaccggattggtctgtgttaccatCCAGCTGAAGCTTCGTACGC
RSC6/X1 (+50r)_R3 ttgctgcacggctcactactcagcaggaggcgtgcaataatgcaCATCGATGAATTCGAGCTCGTT
RHR2/X1pr.re ttgtggtcaaaggcattgcatggtttgt
RHR2/X1orf.fo acaaaccatcgcaatgccttggaccacaaa
RHR2/X1 (-600f)_R3 acgtatggtctgggtgcgaagtaccaccagcttcttcttCATCGATGAATTCGAGCTCGTT
RHR2/X1 (-45f)_X2orf.fo aatatacagaaacaaaaggcaaatatacaaacatcgcaatgCATGGATTGACTACTAAACCTCTATCTTTG
RHR2/X1 (-45f)_F1 atcagaaacaaaaggcaaatatacaaacatcgcaatgCCAGCTGAAGCTTCGTACGC
RHR2/X1 (1r)_X2pr.re agcggcgttgattttcaaagataaaggtttgggtcaaaggcatTCCGAATATTGTTTTTATTGTTTTTATTGTTT
RHR2/X1 (1r)_F1 cgttgattttcaaagataaaggtttgggtcaaaggcatCCAGCTGAAGCTTCGTACGC
RHR2/X1 (+50r)_R3 taaaattatacaaatggctggaatggagggaatcatCATCGATGAATTCGAGCTCGTT
REB1/XmaI_X2re CCGGAcccgggGCTGCTGCTGGTAAGGCTAACAAACAA
REB1/XmaI_X2fo CCGGAcccgggGCATCTGTCACACAACCTGGAACTCA
REB1/XmaI_X1re CCGGAcccgggCCAGTCCTTCCAAGAATGTAGGCAGAAG
REB1/XmaI_X1fo CCGGAcccgggCGCTGAAACAGAAAATGACATTGACGC
REB1/X1pr.re gttatgacctgaaggcatattgattgac
REB1/X1orf.fo gctcaatcaatgctcctcaggtcataac
REB1/X1 (-600f)_R3 aagctcctcgaaaaattgatcagctgatataagcaacaacttCATCGATGAATTCGAGCTCGTT
REB1/X1 (-45f)_X2orf.fo actcagctgttctgctgtaattagaatatttatagctcaatcaatATGGACAGCGTGTCAAACCTTAAGAGC
REB1/X1 (-45f)_F1 actcagctgttctgctgtaattagaatatttatagctcaatcaatCCAGCTGAAGCTTCGTACGC
REB1/X1 (1r)_X2pr.re cacagactcttgatttgcgtttttatcgttatgacctgaaggcatTCTACCTTATCTTTCTTCTTGACGCAG
REB1/X1 (1r)_F1 cacagactcttgatttgcgtttttatcgttatgacctgaaggcatCCAGCTGAAGCTTCGTACGC
REB1/X1 (+50r)_R3 tgttacagggctaaaaaaagataaatgatgaaaaataacgcgtaCATCGATGAATTCGAGCTCGTT
PET9/X1pr.re ggcgttggagacatggctatttgc
PET9/X1orf.fo gcaaatagccatgtcttccaacgcc
PET9/X1 (-600f)_R3 aacagtgcactgccccatgcaattggcgacaagattttaagaCATCGATGAATTCGAGCTCGTT
PET9/X1 (-45f)_X2orf.fo ttatacttcagaatcatacattaacatacatataagcaaatagccATGAGTAGCGACGTAAGCAACAAGAA
PET9/X1 (-45f)_F1 ttatacttcagaatcatacattaacatacatataagcaaatagccCCAGCTGAAGCTTCGTACGC
PET9/X1 (1r)_X2pr.re attaatggcaagtttgttcttctgtttagcgtctactcatTGTCTCAAGGCACAGTATATCACTTCA
PET9/X1 (1r)_F1 atgtcttccaacgccccagtaaaacccccattacctccagccccaCCAGCTGAAGCTTCGTACGC
PET9/X1 (+50r)_R3 acgtaaaagtgttcaaaaaattcatcgttttcttttttatgaCATCGATGAATTCGAGCTCGTT
PDI1/XmaI_X2re CCGGAcccgggGGGAACCTGCAAGAATGGTGGAAAA
PDI1/XmaI_X2fo CCGGAcccgggTCACAGTTCATATAGTCCCGACGGC
PDI1/XmaI_X1re CCGGAcccgggGGAGAGGGAATGGAGGAGGATGAGATAA
PDI1/XmaI_X1fo CCGGAcccgggTGGCTTTGTGTAAGTCGTCGAATGACA
PDI1/X1pr.re ccagcagaaaaacttcataacgggatagatg
PDI1/X1orf.fo catctatcccgttatgaagtttctgctgg
PDI1/X1 (-600f)_R3 fgagttgggcaactgatgggggggttgggggttctggtgatagcggCATCGATGAATTCGAGCTCGTT
PDI1/X1 (-45f)_X2orf.fo aagctgccaagctctacataaagaaaaacatacatctatcccgttATGCAAGTGACCACAAGATTATATCTGCG
PDI1/X1 (-45f)_F1 aagctgccaagctctacataaagaaaaacatacatctatcccgttCCAGCTGAAGCTTCGTACGC
PDI1/X1 (1r)_X2pr.re cagcagggaggaccatgacaggcagcaccagcagaaaaactcatGGTATTATAGCGGGTCTGTT
PDI1/X1 (1r)_F1 cagcagggaggaccatgacaggcagcaccagcagaaaaactcatCCAGCTGAAGCTTCGTACGC
PDI1/X1 (+50r)_R3 aattttaacgtttatctttttatgatttttttaaaaaaacttctCATCGATGAATTCGAGCTCGTT
ITR1/X1pr.re ttgacgtgagataggtatggtatttccattttaacac
ITR1/X1orf.fo gtttaaaatgggaatacacataccatctcactgcaaaa
ITR1/X1 (-600f)_R3 atagcttatacccccaactggttgcgctaaatacgcctCATCGATGAATTCGAGCTCGTT
ITR1/X1 (-45f)_X2orf.fo gtatacatctcttcttcttcttcttctgagtggtttaaATGAAGAATTCGACAGCTGCTAGTTCCAGA
ITR1/X1 (-45f)_F1 catctcttcttcttcttcttcttctgagtggtttaaCCAGCTGAAGCTTCGTACGC
ITR1/X1 (1r)_X2pr.re fgattgcatgcttcttgcagtgagataggtatggtatttccatCTGGAACTAGCAGCTGCGAATCTCTCAT
ITR1/X1 (1r)_F1 gcgatgcttcttgcagtgagataggtatggtatttccatCCAGCTGAAGCTTCGTACGC
ITR1/X1 (+50r)_R3 cgacttaatacaaatgaacttttctatctgttaattCATCGATGAATTCGAGCTCGTT

GUP1/X1pr.re aggatgctgatcagcgacatttcttgcaat
GUP1/X1orf.fo attgcaagaaatgtcgtgatcagcatcct
GUP1/X1 (-600f)_R3 ttccttaatgatacगतatttcgaagtacactgttttgCATCGATGAATTCGAGCTCGTT
GUP1/X1 (-45f)_X2orf.fo aaagcgtgaatcagcatttgttaacagacaattgcaagaaATGTCGATGTTAAGAACTCGGAGTTGCAT
GUP1/X1 (-45f)_F1 aaagcgtgaatcagcatttgttaacagacaattgcaagaaCCAGCTGAAGCTTCGTACGC
GUP1/X1 (1r)_X2pr.re tcggaagtaattaggggagacaggatgctgatcagcgacatCTGACTTGTAGCGATGACGCTCAATTATT
GUP1/X1 (1r)_F1 cggaagtaattaggggagacaggatgctgatcagcgacatCCAGCTGAAGCTTCGTACGC
GUP1/X1 (+50r)_R3 attttcggtttcattttgttactagtagataaaaaaacCATCGATGAATTCGAGCTCGTT
GSP1/X1pr.re ccgtagcagctggggcagacattta
GSP1/X1orf.fo taaatgtctgccccagctgtcaacgg
GSP1/X1 (-600f)_R3 acttttcataगतatcctcctgtcgaagaaaaaacacaacagaaccCATCGATGAATTCGAGCTCGTT
GSP1/X1 (-45f)_X2orf.fo agtccagtacaaaaggaaaatcaaaataaactcatttataataaaATGTCAGCACCTTGCTCAAAACAATGC
GSP1/X1 (-45f)_F1 agtccagtacaaaaggaaaatcaaaataaactcatttataataaaCCAGCTGAAGCTTCGTACGC
GSP1/X1 (1r)_X2pr.re tagttgaaatgttggaacttcaccgtagcagctggggcagacatTGCTGACATTACGAACTTATATAGGTAGTCACA
GSP1/X1 (1r)_F1 tagttgaaatgttggaacttcaccgtagcagctggggcagacatCCAGCTGAAGCTTCGTACGC
GSP1/X1 (+50r)_R3 cataatccacattcaactatttgttatataattaaactctgtttatCATCGATGAATTCGAGCTCGTT
FEN1/X1pr.re gtaacgagtgaaatcaggtctctagacga
FEN1/X1orf.fo tcgtctagaagccatgaattcactcgttac
FEN1/X1 (-600f)_R3 atgctctactctttccatccaatggttggtgaaactCATCGATGAATTCGAGCTCGTT
FEN1/X1 (-45f)_X2orf.fo caaagtgaaggttgaactaattatctcgtctagaagcATGGTAAGTGATTGGAAAAATTTTGCCTC
FEN1/X1 (-45f)_F1 caaagtgaaggttgaactaattatctcgtctagaagcCCAGCTGAAGCTTCGTACGC
FEN1/X1 (1r)_X2pr.re cgaacaacggagcagcatattgagtaacgagtgaaatcactACTTGCTGCAAAATTTAGCTCTCAAGTTCC
FEN1/X1 (1r)_F1 cgaacaacggagcagcatattgagtaacgagtgaaatcactCCAGCTGAAGCTTCGTACGC
FEN1/X1 (+50r)_R3 gtcttttacaataataaatagtgataatagtaacaatCATCGATGAATTCGAGCTCGTT
ERV14/X1pr.re taaataaccaagcaccatttttcttttta
ERV14/X1orf.fo taaaaagaaaaatgggtgctgtgttttta
ERV14/X1 (-600f)_R3 tttctcaataataatataatgtaagtaaccagttgtcaaCATCGATGAATTCGAGCTCGTT
ERV14/X1 (-45f)_X2orf.fo aactgcaattaaagtaaaagtaaaaaaataagaataaaaaagaaaaATGTCAGGAACCCGATTATCGCTATTCG
ERV14/X1 (-45f)_F1 caattaaagtaaaagtaaaaaaataagaataaaaaagaaaaCCAGCTGAAGCTTCGTACGC
ERV14/X1 (1r)_X2pr.re gatacagttgacaacaacagcaaggataaaataaccaagcaccatCTTTACTCTTGCTGCAACTTCTTGCTGTC
ERV14/X1 (1r)_F1 agttgacaacaacagcaaggataaaataaccaagcaccatCCAGCTGAAGCTTCGTACGC
ERV14/X1 (+50r)_R3 tatcttcagttttcctccaactctatggcgcgccataaCATCGATGAATTCGAGCTCGTT
DED1/XmaI_X2re CCGGAcccgggCCCATTTATGATTGGGACAAGAGCAAAA
DED1/XmaI_X2fo CCGGAcccgggCCGCCTCTAGGTTCAAGATGTGGTAATT
DED1/XmaI_X1re CCGGAcccgggGCCAATCAACGGCTAGAGAGAAACAAC
DED1/XmaI_X1fo CCGGAcccgggAGGATTTGCGCTTTGGATGAGG
DED1/X1pr.re cgctcagttcagccataatgaaatgc
DED1/X1orf.fo gcatttcataatattggctgaactgagcg
DED1/X1 (-600f)_R3 tgacaccgattatttaaagctgacacatacगतatatacatgtCATCGATGAATTCGAGCTCGTT
DED1/X1 (-45f)_X2orf.fo ggcaagtggattccgtaagaacaacaagaaaagcatttcataatATGGCAGACTTGCCACAGAAGGTATC
DED1/X1 (-45f)_F1 ggcaagtggattccgtaagaacaacaagaaaagcatttcataatCCAGCTGAAGCTTCGTACGC
DED1/X1 (1r)_X2pr.re gtcggtgatgcttaaaatgttgcactgttcgctcagttcagccatTTTTTCTGTAATACTTTCACTCAAAGGGG
DED1/X1 (1r)_F1 gtcggtgatgcttaaaatgttgcactgttcgctcagttcagccatCCAGCTGAAGCTTCGTACGC
DED1/X1 (+50r)_R3 ttttcaggacttgataaaaaaagtgattggagaagaaaaataCATCGATGAATTCGAGCTCGTT
CLA4/X1pr.re agctgaaagagacataatgttgcactatttc
CLA4/X1orf.fo gaaatagtgcataatgttctctttcagct
CLA4/X1 (-600f)_R3 atatcaaacgctgggtattttccctttccggaaatCATCGATGAATTCGAGCTCGTT
CLA4/X1 (-45f)_X2orf.fo gtaataaatcgaacagtgaaactgaaacataaaaaagaataATGAAGGGCGTAAAAAAGGAAGGTGGATA
CLA4/X1 (-45f)_F1 gtaataaatcgaacagtgaaactgaaacataaaaaagaataCCAGCTGAAGCTTCGTACGC
CLA4/X1 (1r)_X2pr.re gaaatcgttgcagatattcttattcgtcagctgaaagagacatCGAAAGATATTCTTTCGTTTTGCTGATTG
CLA4/X1 (1r)_F1 cgttgcagatattcttattcgtcagctgaaagagacatCCAGCTGAAGCTTCGTACGC
CLA4/X1 (+50r)_R3 aacttttaagtgaaactatttatataatgcccattgCATCGATGAATTCGAGCTCGTT
CDC19/XmaI_X2re CCGGAcccgggCCAGACTGTGCGTAAACTTGACATGGA
CDC19/XmaI_X2fo CCGGAcccgggGGACGCCTTGTGTAACGCTTTTATGA
CDC19/XmaI_X1re CCGGAcccgggGGAAGATTATCTTGCATGGGAGGA
CDC19/XmaI_X1fo CCGGAcccgggAACGGCGGATTCCTCTATGGCGT
CDC19/X1pr.re ccagcaacaacgtttaaaggtcaatctttc

CDC19/X1orf.fo	gaaagattgacctcattaaacgttggtgctgg
CDC19/X1(-600f)_R3	aacggcgggattcctctatggcgtgtgatgtctgtatctgttactCATCGATGAATTCGAGCTCGTT
CDC19/X1(-45f)_X2orf.fo	ttctatttacaagacccaatcaaaacaaataaaacatcatcacaATGCCAGAGTCCAGATTGCAGAGACTAG
CDC19/X1(-45f)_F1	ttctatttacaagacccaatcaaaacaaataaaacatcatcacaCCAGCTGAAGCTTCGTACGC
CDC19/X1(1r)_X2pr.re	agcaacaacgtttaatgaggccaatctttctaatactagacatCGATAGTGCTTTTGTGTAATCTTACAATGG
CDC19/X1(1r)_F1	accagcaacaacgtttaatgaggccaatctttctaatactagacatCCAGCTGAAGCTTCGTACGC
CDC19/X1(+50r)_R3	tttatttgagttgaaaaaaccatgtctttatataaaatttaCATCGATGAATTCGAGCTCGTT
BUL2/X1pr.re	tgaagtgagaatgtaaaagtcattaatac
BUL2/X1orf.fo	gtattaatgacttttacattctccacttca
BUL2/X1(-600f)_R3	tgaattacattcgggctgatatggttgctacataagcCATCGATGAATTCGAGCTCGTT
BUL2/X1(-45f)_X2orf.fo	gcagatttgagatatattctgggaacaaaagaagtattaATGGCCAAGATTTGAACGATTCGGGGTTT
BUL2/X1(-45f)_F1	gcagatttgagatatattctgggaacaaaagaagtattaCCAGCTGAAGCTTCGTACGC
BUL2/X1(1r)_X2pr.re	aggtctcccattttccttgatgaagtgagaatgtaaaagtcattTAGCGTACGATATACCTGTTGACACACAC
BUL2/X1(1r)_F1	tcccattttccttgatgaagtgagaatgtaaaagtcattCCAGCTGAAGCTTCGTACGC
BUL2/X1(+50r)_R3	taataataataatgaatgatattaatagtaaataggtacATCGATGAATTCGAGCTCGTT

ANEXO E

Oligonucleótidos de integración en genoma de construcciones tetO7

Nombre	Secuencia
ERV14/X1 (-600)_F1	cattttctcaataatatatggttaagtaaccagttgtcaaCCAGCTGAAGCTTCGTACGC
ERV14/X1 (around ATG)_R1	ttgacaacaacagcaaggataaataaccaagcaccocatGCATAGGCCACTAGTGGATCTG
TSA1/X1 (-600)_F1	accaggacatatataaaggggaggttaattcgtcagatcaatCCAGCTGAAGCTTCGTACGC
TSA1/X1 (around ATG)_R1	ttaaaagttggagcttgcttttgaacttgagcgaccatGCATAGGCCACTAGTGGATCTG
RHR2/X1 (-600)_F1	agacgtatggctcgggtgcaagtagtaccaccacgtttctttCCAGCTGAAGCTTCGTACGC
RHR2/X1 (around ATG)_R1	ttgattttcaagataaaaggttttgggtgcaagggcatGCATAGGCCACTAGTGGATCTG
GUP1/X1 (-600)_F1	agactgcagtagcctttttttcacgcaacttcagggttcgaCCAGCTGAAGCTTCGTACGC
GUP1/X1 (around ATG)_R1	gaagtaattaggggagacaggatgctgatcagcgacatGCATAGGCCACTAGTGGATCTG
BUL2/X1 (-600)_F1	aatgaatttacattcgggtgatattggttgctacataagcCCAGCTGAAGCTTCGTACGC
BUL2/X1 (around ATG)_R1	ccatttttcccttgatgaagtggaagatgtaaaagtcacGCATAGGCCACTAGTGGATCTG
SDH3/X1 (-600)_F1	tctagaaaattccctcacttatatatatatatatatcGCATAGCTGAAGCTTCGTACGC
SDH3/X1 (around ATG)_R1	gacttattcagtcctaacttgaccatcattgcagacatGCATAGGCCACTAGTGGATCTG
VPS5/X1 (-600)_F1	gcaaaggttatttactataactacgtagtgaacgaaaccaCCAGCTGAAGCTTCGTACGC
VPS5/X1 (around ATG)_R1	tcccaaacaggtgcttctagattatcctcgtagtcacGCATAGGCCACTAGTGGATCTG
ERV14/X2 (-600)_F1	ttgcattatgatagtatagttttgctatacaaaacttcaatCCAGCTGAAGCTTCGTACGC
ERV14/X2 (around ATG)_R1	agaccgcacgaatagcgataatccggttctcgacatGCATAGGCCACTAGTGGATCTG
TSA1/X2 (-600)_F1	taccggagtaatacaaggatcaactatggatgagatttagaCCAGCTGAAGCTTCGTACGC
TSA1/X2 (around ATG)_R1	ttaaatggtgggcttgcttttgaacttctgctaccatGCATAGGCCACTAGTGGATCTG
RHR2/X2 (-600)_F1	gcggcgtttccggtgaagccgtcaagtgaggacttttcgGCAGCTGAAGCTTCGTACGC
RHR2/X2 (around ATG)_R1	ttaaacttcaaaagatagaggttttagtagtcaatccatGCATAGGCCACTAGTGGATCTG
FEN1/X2 (-600)_F1	acggattagtaaacgaaaggacataaaggaaagaaagaaCCAGCTGAAGCTTCGTACGC
FEN1/X2 (around ATG)_R1	gctttctcgaggcaaaaattttccaatcacttaccatGCATAGGCCACTAGTGGATCTG
GUP1/X2 (-600)_F1	tatcgtgcaagcgtgagttcgaaacgctgttttgatgaaCCAGCTGAAGCTTCGTACGC
GUP1/X2 (around ATG)_R1	aaatgtacgatgcaactccagattccttaacatcgacatGCATAGGCCACTAGTGGATCTG
SAC7/X2 (-600)_F1	cttgagggggaggaccagcctcatcgggctaaaactcctCCAGCTGAAGCTTCGTACGC
SAC7/X2 (around ATG)_R1	tcagaagaaggtgtagaaagtagattcatgttaaacatGCATAGGCCACTAGTGGATCTG
BUL2/X2 (-600)_F1	aagagccgcatagtattatggaagaagctcttagcaagCCAGCTGAAGCTTCGTACGC
BUL2/X2 (around ATG)_R1	ttcgggtggaaccccgaatcgttcaaatctttggccatGCATAGGCCACTAGTGGATCTG
ITR1/X2 (-600)_F1	gctcgtcttgccactaaaagtagcaaaatattatgtgaaCCAGCTGAAGCTTCGTACGC
ITR1/X2 (around ATG)_R1	ttcgtccatctggaactagcagctgtcgaattcttccatGCATAGGCCACTAGTGGATCTG
TIF4631/X2 (-600)_F1	gagcattgctgctttttcccgctcctgaccttattaatagCCAGCTGAAGCTTCGTACGC
TIF4631/X2 (around ATG)_R1	tgccgggtgtggggcggtggacctcttgggtcagtcacGCATAGGCCACTAGTGGATCTG
SET3/X2 (-600)_F1	actctgttacgctgaaactaacaatcacacatgtagaggtCCAGCTGAAGCTTCGTACGC
SET3/X2 (around ATG)_R1	ctgatatgacgagaagatagtgattccgggtgaagtcacGCATAGGCCACTAGTGGATCTG
CLA4/X2 (-600)_F1	taccgtatacgtttgatttctcagccccttggaaagacacCCAGCTGAAGCTTCGTACGC
CLA4/X2 (around ATG)_R1	ttataagatatccatccttcttttttaagcccttcatGCATAGGCCACTAGTGGATCTG
VPS5/X2 (-600)_F1	gtgtagtggaaagcaaaatgcactgcgcttatgtgaaaaagCCAGCTGAAGCTTCGTACGC
VPS5/X2 (around ATG)_R1	atgcttgggctcgcgatgtgaecgcttctcgttctccatGCATAGGCCACTAGTGGATCTG
SDH3/X2 (-600)_F1	agatataagagtcctcataatggaatcggattttctacatCCAGCTGAAGCTTCGTACGC
SDH3/X2 (around ATG)_R1	ccaaatacagatggttactccttgaatggttgctttcGCATAGGCCACTAGTGGATCTG

ANEXO F

Artículo 1: aceptado en BMC Molecular biology and evolution

Increased rates of protein evolution and asymmetric deceleration after the whole-genome duplication in yeasts

Diana Ascencio¹, Soledad Ochoa¹, Luis Delaye² & Alexander DeLuna¹

¹Unidad de Genómica Avanzada (Langebio), Centro de Investigación y de Estudios Avanzados del IPN, Irapuato, Guanajuato, Mexico

²Departamento de Ingeniería Genética, Centro de Investigación y de Estudios Avanzados del IPN, Irapuato, Guanajuato, Mexico

Diana Ascencio (diana.ascencio@cinvestav.mx)

Soledad Ochoa (msoledad@lcg.unam.mx)

Luis Delaye (luis.delaye@cinvestav.mx)

Correspondence to:

Alexander DeLuna (alexander.deluna@cinvestav.mx)

Abstract

Background: Whole-genome duplication (WGD) events have shaped the genomes of eukaryotic organisms. Relaxed selection after duplication along with inherent functional constraints are thought to determine the fate of the paralogs and, ultimately, the evolution of gene function. Here, we investigated the rate of protein evolution (as measured by dN/dS ratios) before and after the WGD in the hemiascomycete yeasts, and the way in which changes in such rates relate to molecular and biological function.

Results: For most groups of orthologous genes (81%) we observed a change in the rates of evolution after genome duplication. Genes with atypically-low dN/dS ratio before the WGD were prone to increase their rates of evolution after duplication. Importantly, the paralogs were often different in their rates of evolution after the WGD (50% cases), however, this was more consistent with an asymmetric deceleration in the protein-evolution rates, rather than an asymmetric increase of the initial rates. Functional-category analysis showed that regulatory proteins such as protein kinases and transcription factors were enriched in genes that increase their rates of evolution after the WGD. While changes in the rate of protein-sequence evolution were associated to protein abundance, content of disordered regions, and contribution to fitness, these features were an attribute of specific functional classes.

Conclusions: Our results indicate that strong purifying selection in ancestral pre-duplication sequences is a strong predictor of increased rates after the duplication in yeasts and that asymmetry in evolution rate is established during the deceleration phase. In addition, changes in the rates at which paralogous sequences evolve before and after WGD are different for specific protein functions; increased rates of protein evolution after duplication occur preferentially in specific protein functions.

Keywords: Gene duplication; whole-genome duplication; protein-evolution rate; purifying selection; asymmetric evolution; phylogenetic analysis by maximum likelihood; biological function; *Saccharomyces cerevisiae*

Background

Analyses of genomic sequences indicate that whole-genome duplication (WGD) events have had strong influence in the genomes and evolution of several eukaryotic lineages of vertebrates [1], flowering plants [2], and fungi [3, 4]. Compared to duplication events that occur at smaller scales, WGD increases the potential of functional and morphological innovation by increasing the rate of duplicate retention of specific types of genes [5]. In particular, the retention of dosage-sensitive paralogs can be favored, leading to regulators and subunits of protein complexes in more than one copy [6, 7]. Likewise, regulatory and developmental genes have evolved by WGD. It has been documented that genome duplications are responsible for 90% of the increase in regulatory genes in *Arabidopsis thaliana* [8, 9]. In vertebrates, homeotic genes and multiple receptors have expanded preferentially by means of WGD [10, 11]. Therefore, the great morphological complexity and diversity in eukaryotes is, to a large extent, an effect of the amplification of the regulatory repertoire by genome duplication.

The budding yeast *Saccharomyces cerevisiae* evolved from an ancestor which underwent a WGD event, dated ~100 mya [3, 4]. This genome duplication in the hemiascomycete yeasts is perhaps the best documented event of its kind. Recent phylogenetic analysis strongly suggest that the baker's yeast genome is the result of an interspecies hybridization followed by WGD to restore fertility [12, 13]. Almost

10% of the genes were retained in two copies after the genome duplication, giving rise to more than 500 extant paralogous gene-pairs [14]. Even though only a small proportion of genes were retained in duplicate, the WGD had an impact on the lifestyle and metabolism of yeast species. For instance, it has been suggested that gene duplication improved the glycolytic flux in *S. cerevisiae*, which may have led to the colonization of high-sugar content ecological niches [15, 16].

There are biological functions associated to genes that are more likely to be retained in two copies after gene or genome duplication. It has been documented that the WGD in *S. cerevisiae* favored the retention of ribosomal proteins, enzymes of carbohydrate metabolism, and signal-transduction kinases in duplicate [17]. Interestingly, the functional classes of paralogs retained after WGD are similar across diverse taxa ranging from yeast to *Arabidopsis thaliana* and *Paramecium* spp. [18]. Recurrent gene conversion events [19] can favor the maintenance of identity between duplicated copies of genes that code for highly constrained proteins, such as ribosomal subunits and histones [20, 21]. Still, even though WGD paralogs tend to retain an important degree of functional overlap and compensate for each other's loss [22, 23], many instances of asymmetry in evolutionary rates and gene loss indicate extensive neofunctionalization after the WGD in yeast [24]. Importantly, the interspecies hybridization that gave rise to the *Saccharomyces* lineage could also have

had strong implications on which genes were conserved in two copies after the WGD, as sequences were not identical at the beginning [12].

The action of natural selection before and after gene duplication has been analyzed in a systematic manner for several model organisms. These studies have shown that the rates of evolution tend to increase after gene duplication due to a relaxation of purifying selection [25-28]. Gene duplication is followed by a brief period of relaxed selection that declines rapidly, but decelerated evolutionary rates rarely revert to ancestral rates [25, 28, 29]. In addition, asymmetric rates of protein evolution between the duplicates has been documented for different eukaryotes [28, 30-32]. However, it is unclear when is this pattern of protein evolution established and to what extent it holds for different functional classes of genes that are retained in two copies after a WGD event.

In this study, we asked whether the rates of evolution change at different points in the evolutionary history of a large set of paralog pairs of the hemiascomycetes fungi, paying special attention to the rates before and after the WGD. We also asked how molecular function and other gene characteristics are associated to such evolution-rate dynamics. To this end, we took advantage of the Yeast Genome Order Browser (YGOB) compendium of sets of orthologous genes, which takes into account the phylogeny of 20 Saccharomycetaceae genomes [14]. With such catalogue, we tested different

evolution models with a phylogenetic analysis by maximum-likelihood (PAML) approach. For most orthogroups, the rates of evolution increased immediately after the WGD event in a symmetric manner and, in many cases, this was followed by asymmetric deceleration of the evolutionary rates. In addition, we observed that the specific patterns of evolution depended on the molecular function of the duplicate genes, regardless of gene expression, disordered content, or contribution to fitness.

Results

Ancestral proteins evolving at very slow rates undergo relaxation of purifying selection after the WGD

The classic model of evolution by gene duplication suggests that the co-occurrence of two paralogs that can compensate for each other's mutations leads to increased rates in sequence divergence after duplication [33, 34]. We asked to what extent this scenario is consistent with the dN/dS ratios estimated for phylogenetic lineages that diverged before or after the WGD in the hemiascomycetes. We used a maximum-likelihood approach to test the hypothesis of different average rates of evolution between non-WGD (ω_0) and post-WGD (ω_{wgd}) branches against the null hypothesis of sequences evolving at equal rates across the phylogeny (see example in **Figure S1**). Such tests were performed for 535 orthologous groups of genes (orthogroups) from 20 yeast species [14] in which the

two WGD-paralogs (ohnologs) have been maintained in *S. cerevisiae*. Such orthogroups include both ohnologous clades and their co-orthologs in non-WGD species.

We found that the rates of evolution are significantly different between non-WGD and post-WGD sequences for a large fraction (81%) of the analyzed orthogroups (**Figure 1A**, " ω -change set", orange bars). The remaining orthogroups showed low likelihood of change in rate of evolution after WGD, thus the hypothesis of constant rates across the phylogeny could not be rejected (**Figure 1A**; " ω -constant set", blue bars). The average dN/dS ratios were always less than 1.0 for both non- and post-WGD branches, but were typically higher in post-WGD sequences ($p < 10^{-60}$, Wilcoxon rank-sum test), indicating a general trend of relaxation of purifying selection after the WGD event (**Figure 1B**). We repeated this analyses after removing all orthogroups corresponding to ribosomal proteins, since such genes are subject to recurrent gene-conversion events after duplication [20]. A slightly larger fraction (86%) of 480 such orthogroups showed a change in evolutionary rate after duplication, and the median dN/dS ratio was also significantly higher in post-WGD branches ($p < 10^{-71}$), indicating that ribosomal-protein genes were not a major driver of the observed trends. Together, these results confirm that most WGD genes increase their dN/dS ratio after duplication.

To further describe the changes in the rates of evolution, we compared the average dN/dS ratios of non-WGD and post-WGD sequences within each orthogroup. A vast majority of orthogroups fell in or above the diagonal; only a minor fraction showed the opposite scenario in which the rates of evolution decreased after duplication (**Figure 1C**). We found that the overall distributions of evolutionary rates were different for branches before and after the WGD, with post-duplication sequences showing a more uniform distribution of rates centered at around dN/dS=0.1; non-WGD sequences showed a wider distribution of their evolutionary rates. Importantly, we observed that yeast orthogroups could be readily divided in two groups based on their pre-duplication dN/dS rates (**Figure 1C**, yellow/purple histogram bars). The null hypothesis of equal evolution rates was more likely to be rejected in the orthogroups with very low pre-duplication dN/dS rates ($p < 10^{-8}$, Fisher's exact test); the difference was also significant after excluding ribosomal-protein orthogroups ($p < 10^{-5}$). Indeed, virtually all orthogroups (98%) under such strong purifying selection before duplication changed their rates of evolution after the WGD event in the Saccharomycetaceae family (**Figure 1D**).

To test for sensitivity in sampling, we carried out our analysis using a different data set of groups of orthologous genes in yeasts, namely from the Fungal Orthogroups Repository (FOR) [35]. This data set includes sequences from species which diverged from each other much

before the WGD in the baker's yeast lineage, such as those from the *Candida*, *Yarrowia*, *Aspergillus*, *Neurospora*, and *Schizosaccharomyces* genera. A high likelihood of change in the rates of protein evolution after duplication was obtained for most of the orthogroups that were represented in both data sets (**Figure S2**, lime dots in panel A). Overall, 70.8% (274 out of 387) of the ω -change orthogroups from either data set showed a significant change in the rates of protein evolution in both analyses (**Figure S2**, panel B). While the dN/dS ratios were similar for post-WGD sequences in both data bases, sequences from non-WGD sequences showed important differences (**Figure S2**, panel C). This was expected given that the YGOB and FOR databases differ mostly in terms of sampling from species that diverged before the WGD.

Many orthogroups show asymmetric evolution after duplication, which is consistent with differences in deceleration rather than in acceleration rates

To examine whether the rates of paralogous-protein evolution have changed at different points of the yeasts phylogeny, we generated a set of new hypothesis tests (**Figure 2A-E**). First, we confirmed that the WGD was followed by a brief period of accelerated protein-evolution [25, 28, 29]. Specifically, the median dN/dS rate of immediate post-WGD was higher than that of ancestral branches ($p < 10^{-108}$, Wilcoxon rank-sum test), while modern post-WGD had a lower median rate than that of immediate post-

WGD branches, but still higher than the typical non-WGD rate ($p < 10^{-49}$) (**Figure 2B**). In terms of individual cases of orthologous proteins, we found that 320 (59.8%) orthogroups were consistent with a scenario of different non-WGD, immediate post-WGD, and modern post-WGD branches ($\omega_0 \neq \omega_{a1,b1} \neq \omega_{a2,b2}$; **Figure 2C**), which again confirms that increased rates of evolution immediately after WGD are in many cases followed by a deceleration phase.

Paralogs frequently evolve asymmetrically after duplication. We therefore asked whether asymmetric changes in the rates of evolution could better explain the natural history of fungal orthogroups. To this end, we first tested the hypothesis of different acceleration rates, namely different evolutionary rates between the paralogous in immediate post-WGD branches. Intriguingly, the null hypothesis of equal rates between the two paralogous branches ($\omega_{a1} = \omega_{b1}$) could not be rejected in virtually any of the orthogroups (**Figure 2D**). In contrast, the null hypothesis of equal rates between the entire post-WGD clade ($\omega_{a1,a2} = \omega_{b1,b2}$) was rejected for a considerable fraction (50.5%) of cases (**Figure 2E**), consistent with previous observations of asymmetric evolution of paralogs. These results suggest that asymmetric evolution is established during the deceleration phase (return of constraint), rather than during the initial stage of increase in protein-evolution rate (relaxation of constraint).

To ask whether asymmetric rates were linked to strong purifying selection in non-WGD branches, we compared the proportion of cases in which the paralogous clades showed significantly-different rates, both for orthogroups with strong non-WGD purifying selection and those with weak-to-moderate purifying selection (as previously defined in **Figure 1B,C**). We found no significant difference in the numbers of genes in such classes ($p=0.61$, Fisher's exact test), indicating that a strong purifying selection before duplication is a good predictor of rate-change after duplication, but not of whether such changes are symmetric or asymmetric.

Different protein functions show contrasting patterns of sequence evolution

Following the WGD, many cytosolic ribosomal proteins, enzymes (specially carbohydrate metabolism), and regulatory proteins (mostly kinases) were retained in two copies in *S. cerevisiae* [17]. Indeed, ~39% of the orthogroups in our analysis are related to such three protein functions (71 regulatory proteins, 82 enzymes, and 55 ribosomal proteins). Using this set of highly-retained duplicates, we asked whether proteins with different biochemical roles display similar or disparate patterns of rate-change dynamics. Our results showed that the fraction of ω -change orthogroups depended on protein function. For instance, duplicated regulatory genes (protein kinases and phosphatases, transcription factors, and chromatin

remodelers) were enriched in ω -change orthogroups, while ribosomal proteins were depleted in such scenario (**Figure 3A**). Regulators were also enriched in ω -change orthogroups after removing all ribosomal-protein genes from the analysis ($p=0.001$). Metabolic enzymes showed similar trends of changes in evolution rate as the complete set of orthogroups; the same pattern was observed when focusing only on 32 enzymes with carbohydrate-metabolism role (data not shown).

While a general increase of the typical dN/dS ratio was observed after duplication in the "all orthogroups" category (see **Figure 2B**), regulatory proteins such as protein kinases were a special case in which this trend was exacerbated; median non-WGD and post-WGD dN/dS ratios were significantly different between each other ($\omega_0 \neq \omega_{a,b}$; **Figure 3B**; $p < 10^{-15}$, Wilcoxon rank-sum test). The median ratios were also different in the cases of metabolic enzymes ($p < 10^{-5}$) and ribosomal proteins ($p = 0.007$). In fact, the typical rates of regulatory proteins were marginally higher than all orthogroups only in post-duplication branches, while this was not the case for metabolic enzymes (black dots and black lines in **Figure 3B**). Interestingly, ribosomal proteins evolved much faster than other genes in nonWGD branches, but appeared to evolve slower after gene duplication. This is most likely do to recurrent gene conversion events that have taken place in this set of highly constrained genes [20]. Taken together, these results suggest that biochemical function determines the

patterns of sequence evolution of genes that are retained after long evolutionary times following WGD. In particular, the acceleration of protein-evolution rate is specific to certain biological functions, such as regulatory proteins.

Protein abundance, disorder content, and contribution to fitness are determinants of the change in rate of evolution after WGD, but are strongly associated to function

The rate of coding-sequence evolution depends on the level of functional constraint of the protein. In particular, gene expression is an important determinant of the rate of protein evolution, whereby highly abundant proteins evolve at slower rates [36, 37]. Selection against errors in translation, protein folding, or protein misinteraction are thought to explain this trend [38]. We therefore asked to what extent protein abundance could explain the changes in the rates of evolution after duplication and how it relates to protein function. To this end, we used quantitative data of absolute GFP-tagged protein abundances in *S. cerevisiae* [39]. We compared the distributions of protein abundances between individual paralogs in ω -change and ω -constant orthogroups (**Figure 4A**). In agreement with previous inferences, we confirmed that the ω -constant proteins are typically more abundant in *S. cerevisiae* than the ω -change paralogs ($p < 10^{-10}$, Wilcoxon rank-sum test, median protein abundance levels were 155 and 70.9 ppm for ω -constant and ω -change paralogs, respectively). Different protein functions

were associated to important differences in protein abundance, as expected (*e.g.* ribosomal proteins were more abundantly expressed than regulatory proteins). However, the general trend of higher protein abundance of the ω -constant versus the ω -change set was not evident and not significant within different protein functions (**Figure 4A**; $p=0.43$, $p=0.65$, and $p=0.21$ for regulators, enzymes, and ribosomal proteins, respectively). It must be noted, however, that the small sample size in certain cases of functional groups may account for the lack of statistical significance in the differences of protein abundance between the ω -constant and ω -change sets.

To gain further insight on the functional constraints associated to the change in evolutionary rates after duplication, we also considered the relation between changes in rate of evolution and content of disordered amino acids. While many proteins need to adopt a well-defined structure to carry out their function, a large fraction of the proteome consists of polypeptide segments that are not likely to form a defined three-dimensional structure, but are nevertheless functional [40]. Such disordered proteins have been observed to evolve at high rates [41]. Indeed, we confirmed that the ω -change set of WGD paralogs was associated to *S. cerevisiae* proteins with a higher content of disordered amino acids, compared to the ω -constant set (**Figure 4B**; $p < 10^{-11}$, Wilcoxon rank-sum test). The median number of amino acids in disordered regions was 55 and 144 for ω -constant and

ω -change proteins, respectively. However, when looking at individual functions, no difference was observed in the median content of disordered amino acids of the ω -constant and ω -change gene sets ($p=0.13$, $p=0.38$, and $p=0.46$ for regulators, enzymes and ribosomal proteins, respectively). Important differences were observed among functional groups in terms of disorder content: while regulatory proteins had a significantly higher median content of amino acids in disordered regions than the entire set of orthogroups ($p<10^{-10}$), both metabolic enzymes and ribosomal proteins had typically a lower content ($p<10^{-5}$ and $p<10^{-7}$, respectively). Thus, the number of amino acids in disordered regions was more a property of the functional category and was not directly associated with evolution-rate changes after the WGD.

Essential and non- dispensable genes typically evolve at slower rates than dispensable genes [36, 42]. However, gene duplication may hide the fitness cost of single-gene deletions due to the functional compensation provided by the functionally-overlapping paralogous gene. Indeed, the fitness cost of deleting single WGD genes is usually small [22]. We took advantage of available genome-wide high-resolution fitness data of single-gene knockouts in *S. cerevisiae*. Consistent with previous observations, we found that ω -constant orthogroups had individual paralogs with a higher contribution to fitness (lower fitness of gene deletion) than paralogs in the ω -change set (**Figure 4C**; $p<10^{-3}$, Wilcoxon rank-sum test).

However, once again this depended more on the functional group; no difference was observed within each of the functional groups ($p=0.11$, $p=0.51$, and $p=0.90$ for regulators, enzymes and ribosomal proteins, respectively). As expected, ribosomal proteins usually had a strong contribution to fitness in *S. cerevisiae* and maintained a constant evolutionary rate in non-WGD and post-WGD branches of the phylogeny. Actually, we observed no significant difference between ω -constant and ω -change orthogroups after removing all ribosomal proteins from the analysis ($p=0.39$), indicating that the association between rate-change and fitness contribution is driven primarily by ribosomal-protein genes, which have a strong contribution to fitness and tend to show no change in rate due to gene conversion. Together, these results suggest that an interplay between the inherent gene-function 'dispensability' and the degree of functional compensation and divergence between each paralog pair determines the relationship between fitness cost of gene deletion and change in rate of evolution after WGD.

Together, these results confirm that changes in the rate of protein sequence evolution after WGD are associated to protein features such as abundance, content of disordered regions, or contribution to fitness, but are strongly connected to protein function.

Functional analysis of duplicate pairs with asymmetric evolution

Finally, we asked whether asymmetric paralogous evolution was more common in certain protein functions. We observed that regulators and metabolic enzymes had a higher fraction of paralogs with asymmetric rates of evolution, but these differences were not significant (**Figure 5A**). Ribosomal proteins were significantly depleted in such scenarios of asymmetric rates after the WGD, as expected given their structural and functional constraints and, more importantly, recurrent gene-conversion events that have taken place in this set of genes after duplication.

Differences in protein abundance between the paralogs may explain asymmetry in the rates of evolution. To test this idea, we used data from *S. cerevisiae* to compare the abundances of proteins encoded by the slow and fast-evolving paralogs (only within the 270 orthogroups that showed asymmetric rates of evolution between the two paralogous clades in test *R4*). While some of the slow-evolving paralogs were relatively more abundant, we observed that the distribution of relative expression levels was mostly centered around one (*i.e.* both paralogs were expressed at similar levels) and that in some cases the paralog from the high-rate clade was actually present at higher relative levels (**Figure 5B,C**).

Focusing on different functional annotations, we found that metabolic enzymes were not consistent with the expected trend: only in three cases the more abundant *S. cerevisiae* paralogous

protein belonged to the low-rate clade (stars in **Figure 5C**). Most of this trend could be explained by strong differences in codon usage bias, suggesting that expression change followed by codon-usage change may be the dominant force acting on several genes encoding for metabolic enzymes. Meanwhile, regulatory proteins were usually expressed at low levels in both paralogs (data not shown); where some difference in relative-protein abundance was observed, the bias was either towards the low or fast-rate paralog. Hence, differences in expression were not enough to explain the asymmetric rates within protein kinases and transcription factors. In conclusion, the fact that highly-abundant proteins evolve at lower rates than their less-abundant WGD paralogs holds only for specific cases of duplicate genes.

Discussion

We examined the changes in evolutionary rates of duplicate-gene sequences after the WGD in the hemiascomycetes yeasts, using a maximum-likelihood approach within individual groups of orthologous genes. We took advantage of the highly curated set of WGD paralogs provided by the Yeast Genome Order Browser (YGOB) project [14]. In doing so, our study expands on previous work by Scannell and Wolfe [28], who used amino-acid evolution models on concatenated sequences from eight yeast species to describe the changes protein sequence evolution and asymmetric evolution following WGD in the baker's yeast

lineage. Moreover, our approach allows to interrogate the protein-evolution dynamics at different points in the natural history and for different types of proteins, providing further insight on the impact of WGD in the evolution of gene function.

Our results indicated that a significant increase in the dN/dS ratio occurred after the WGD in most orthogroups in which both paralogs were maintained. This trend of accelerated protein evolution after gene duplication has been consistently observed in yeast and other eukaryotes for different groups of paralogs and their co-orthologues [25-28]. Interestingly, we identified a group of proteins under strong purifying selection, which in almost all cases increased their rates of evolution after duplication. This, together with the fact that conserved proteins are more likely to be retained in duplicate [27], suggests that the WGD had a global effect on the proteome through a general relaxation of purifying selection in a considerable set of highly conserved proteins.

More than 50% of the orthogroups showed asymmetric rates of protein evolution between the two paralogs, consistent with previous inferences for yeast and other eukaryotes [28, 30-32]. However, our detailed evolutionary analysis in which we interrogated the evolutionary rates at different points of the phylogeny suggested that overall rate asymmetry is established during the modern post-WGD deceleration phase (return to constraint), rather than in the initial phase of relaxation

of constraint following the WGD (see model in **Figure 6**). In terms of duplicated-gene function, such an asymmetric deceleration of protein-evolution rate is consistent with a subneofunctionalization scenario in which rapid subfunctionalization during the initial relaxation of purifying selection is followed by prolonged neofunctionalization during the asymmetric-deceleration phase [43].

In addition, we found that proteins with different functions can show contrasting patterns of sequence evolution. Protein kinases, transcription factors, and other regulatory proteins were very likely to change their rates of sequence evolution after the WGD. Accordingly, accelerated evolution rates in kinases and regulatory proteins has been previously inferred by comparing *S. cerevisiae* paralogs with their non-WGD *Kluyveromyces waltii* co-orthologues [4]. In addition, there is recent evidence of signal-transduction pathway expansion by means of genome duplication in *Phycomyces blakesleeanus* and other Mucorales fungi [44]. These observations suggest that signal-transduction pathways may have experienced important functional changes after the WGD, given that many protein kinases evolved slowly in ancestral branches and increased significantly after duplication. For example, the *S. cerevisiae* paralogs *DBF2* and *DBF20*, and their *Candida albicans* co-orthologue *CaDBF2* belong to an orthogroup with a strong and significant change in evolutionary rates after duplication (**Table S1**). The essential

CaDBF2 gene from *C. albicans* codes for a single protein with a role in cytokinesis and mitotic spindle organization [45]. In *S. cerevisiae*, *Dbf2* and *Dbf20* are serine/threonine kinases with a role in cytokinesis [46]. While these paralogs have partially overlapping roles, there is experimental evidence of functional divergence: *DBF2* is expressed at a higher level than its paralog and also has a more prominent role in mitosis compared to *Dbf20* [46, 47]. Interestingly, *Dbf2* has also been shown to play a role in *S. cerevisiae* regulating the stability of mRNAs by co-transcriptional binding [48].

Transcription factors were among the proteins that were more often subject to strong purifying selection before the WGD in yeasts (**Figure 3**). Gene duplication plays an important role in the evolution and functional diversification of transcription factors; the great morphological complexity in plants and animals could be a consequence of the amplification of the regulatory repertoire by WGD [10,11]. Changes in the amino acid sequences of such regulatory proteins were limited by their functions and that such strong constraint was relieved after gene duplication. In a genome duplication context, this could have had important implications in the systems-level evolution of regulatory pathways. Indeed, it has been shown that WGD lead to widespread partitioning of expression networks in yeast [49].

Our results showed that genes with significant changes in the rates of protein evolution after WGD share some features. Proteins in ω -change orthogroups were less abundant, had higher content of intrinsic disorder, and were more dispensable in *S. cerevisiae* than ω -constant orthogroups (**Figure 4**). However, such features were intrinsically associated to molecular function, rather than to changes in the rates of evolution after the WGD. For instance, orthogroups of ribosomal proteins are highly expressed and have a stronger contribution to fitness, while protein kinases and other regulators tend to have higher intrinsic-disorder content. Indeed, it has been suggested that protein function sets important constraints in the way in which protein sequences change with time [42]. It must be noted that an important limitation is that most data has been collected in the budding yeast *S. cerevisiae* and less experimental data is available across other yeasts across the phylogeny. It would be interesting to analyze, for instance, whether proteins in orthogroups with atypical slow rates before WGD, such as kinases, are more abundant--and thus evolved slower--in species that diverged from *S. cerevisiae* before the WGD. Asymmetric evolution was rarely observed for ribosomal proteins, which were clearly depleted in orthogroups in which the evolution rates were significantly different among pre-duplication and both post-duplication branches. This could be readily explained by recurrent gene-conversion events after duplication in WGD paralogs coding for cytosolic ribosomal proteins [20].

Finally, we found that differences in expression level, based on available data for *S. cerevisiae*, were not always associated to the asymmetric rate of evolution. We did find both cases of orthogroups in which the more abundant protein evolved either at a significantly slower or faster rate compared its less abundant paralog. Among the most evident cases of the latter were the Cdc19 and Pyk2 pyruvate kinases, which catalyze the conversion of phosphoenolpyruvate to pyruvate in the final step of glycolysis. *CDC19* is essential for growth on glucose or other fermentable sugars as the sole carbon source, but are dispensable for growth on ethanol or lactate, indicating that there is an alternate route for pyruvate synthesis [50]. Overexpression of *PYK2* restores growth on glucose in *cdc19* mutant cells [51]. High Cdc19 levels are due to activation by fructose-1,6-bisphosphate (FBP), while the *PYK2* paralog is subject to glucose repression and appears to be insensitive to FBP levels, suggesting that it is active when FBP levels are too low to activate *CDC19* [51]. The phosphogluconate dehydrogenase isoenzymes are a similar instance of the slower-evolving paralog expressed at a higher level [52]. *GND1* is the major isoform, is induced during growth on D-glucono-delta-lactone, and is repressed during growth on ethanol or lactic acid [52]. In both *CDC19/PYK2* and *GND1/GND2* cases, enzymes of the paralogous pairs have similar biochemical roles, but are expressed in contrasting patterns and display strong differences in

their codon-adaptation biases. This was also the case for other metabolic-gene pairs with strong evolutionary-rate and expression asymmetry, namely *GDH1/GDH3*, *GPD1/GPD2*, *TRR1/TRR2*, and *ADE16/ADE17*. Changes in gene expression followed by changes in codon usage are likely the dominant evolutionary force acting on several metabolic genes.

Conclusions

Our results indicate that the most prominent cases of changes in protein evolution rate after the WGD in yeast occurred in proteins that were subject to strong purifying selection before duplication, which experienced a brief increase in protein-sequence evolution [28] followed by an asymmetric deceleration in their rates of evolution. An asymmetric return to evolutionary constraint could lead to important functional differences among proteins from species that diverged after the WGD; characterizing the orthologues of gene duplicates or reconstructing ancestral proteins [53] will provide valuable information to understand the connection between sequence and functional divergence of duplicate genes. In addition, we found that the way in which protein-evolution rates changed after the WGD in yeasts is different for different protein functions. Importantly, it remains to be analyzed to what extent and how the inter-species hybridization event that likely preceded the WGD in yeasts [12] may have impacted the evolutionary-rate

dynamics and functions of WGD paralogs; many paralogous branches may have actually begun as non-identical orthologues. It can be anticipated that focusing on specific functional groups across the yeasts phylogeny will provide valuable information to understand the patterns of evolution of paralogous proteins in the context of genome duplication.

Methods

Acquisition of sequences, alignments, and phylogenetic trees. We obtained the protein amino acid and DNA sequences from the Yeast Genome Order Browser (YGOB, <http://ygob.ucd.ie/>), which contains sequences from 20 hemiascomycete species: twelve species diverged from one another after the WGD and eight diverged before the WGD [14]. Distantly-related species such as *Candida albicans* and *Schizosaccharomyces pombe* are excluded from the database due to low gene-order conservation. The YGOB database is a manually curated database that uses both sequence similarity and genome context (synteny) to assess the homology relationships of genes. It provides groups of ohnologs and their co-orthologs (pillars) from all species; here we refer to such pillars as orthogroups (groups of orthologs). To our ends, we selected orthogroups which contain the two WGD paralogous clades with the two gene copies retained at least in *S. cerevisiae* (ohnologs) and their pre-duplication co-orthologues from non-

WGD species in the phylogeny (see example in **Figure S1**).

Starting from 546 orthogroups, groups with less than 13 sequences and with less than three nonWGD sequences were discarded; the remaining 535 orthogroups were used for further analysis. Individual files containing the DNA and amino acid sequences from each orthogroup were organized and the stop codons from DNA sequences were trimmed and compared with correspondent amino acid sequences. The amino acid sequences of each orthogroup were aligned with MUSCLE 3.6 [54]. Nucleotide alignments with PAML format were generated using the amino-acid alignments as a template with PAL2NAL [55]. Finally, the phylogenetic trees were generated based on the topology of the species phylogeny retrieved from the YGOB database and adapted to each orthogroup using the ETE3 toolkit, which allows the manipulation of tree topology by rooting, pruning, or sorting branches [56].

In addition, we used sequences from an alternative data set, the Fungal Orthogroups Repository (<http://www.broadinstitute.org/regev/orthogroups/>), which contains sequences from 23 ascomycete species. In this case, the data set includes distant related species such as *S. pombe* [35]. This non-supervised database was generated using the SYNERGY algorithm [57], which uses sequence similarity and a given species phylogeny to reconstruct the evolutionary history of all genes in a large group of species. As a result, gene trees that reconstruct the evolutionary history of the genes, including gene duplications and losses, are obtained. Here, each gene orthogroup contains all of the orthologous genes present

in species of the database [35]. Again, we selected orthogroups which contain the two WGD paralogous clades with the two gene copies retained at least in *S. cerevisiae*. From 545 orthogroups, we removed 83 based on sequence quality (frequency of undetermined nucleotides) and quantity (at least six gene sequences), and took the remaining 462 orthogroups for further analysis. We generated the alignments as described above for sequences from the YGOB database.

Phylogenetic Analysis by Maximum Likelihood with CodeML. Once we had all the curated alignments and the corresponding phylogenetic trees, we contrasted in each case two evolutionary models using CodeML, a package from PAML 4.4 [58]. To test for differences in the evolution rates before and after WGD, in a first hypothesis test we identified the WGD in each phylogeny based on the topology of the tree and grouped the branches in non-WGD (ω_0) and post-WGD (ω_{wgd}) sequences. Then, we used these topologies, alignments, and trees as CodeML input using a standard control script. As an alternative model (*R1*), we used the user-specified “branch” model which allows the calculation of a distinct ratio for the ω_0 and ω_{wgd} branches (Codonfreq=2). This alternative model was compared with the null model in which all the branches have the same dN/dS ratio. The likelihood ratio test was used to compare the null and alternative models and was calculated as $2(L_1 - L_0)$, where L_1 is the maximum likelihood value of the alternative hypothesis and L_0 the maximum likelihood value of the null

hypothesis. A χ^2 distribution with one degree of freedom was used to calculate the p -value; null hypotheses were rejected with a $p < 0.05$ cutoff (see example in **Figure S1**). Using this same approach, we generated different pair of null-alternative hypotheses (*R2*, *R3*, and *R4*, see **Figure 2**) to test for differences in the rates of evolution at different points in each of the orthogroups.

Protein functional groups, abundance, fitness, and intrinsic disorder data.

Functional categories were manually defined based on the functional annotations for the *S. cerevisiae* paralogs obtained from the Saccharomyces Genome Database (<http://www.yeastgenome.org>). The regulatory function includes genes annotated as protein kinases, phosphatases, transcription factors, and proteins involved in chromatin remodeling. Metabolic function included enzymes involved on biosynthetic and catabolic processes, while ribosomal function included proteins which are structural components of the ribosome. In addition, we used the protein abundance data based on spectral counting of strains expressing GFP-tagged proteins [39]. Fitness data was obtained from high-resolution competition-based experiments describing the fitness of single-gene deletions relative to wild-type cells [59]. Finally, intrinsic disorder content data was from Landry et al. [60], where intrinsically unstructured regions were predicted using the DISOPRED algorithm [61].

Table S1 includes all dN/dS data, likelihood-ratio tests' results, and *S. cerevisiae* protein properties and functions for the 535

orthogroups of the YGOB database, while **Table S2** includes results for 462 orthogroups using the FOR database.

List of abbreviations

LRT, Likelihood-Ratio Test

PAML, Phylogenetic Analysis by Maximum Likelihood

WGD, Whole-Genome Duplication

Declarations

Ethics approval and consent to participate: Not applicable.

Consent for publication: Not applicable.

Availability of data and materials: The raw datasets generated in the current study are available at <http://www.langebio.cinvestav.mx/deluna/WGDO.rar>

Competing interests: The authors declare that they have no competing interests.

Funding: This work was funded by the Consejo Nacional de Ciencia y Tecnología de México (CONACYT grant CB2011/164889 to AD). DA had a doctoral scholarship from CONACYT (267799).

Authors' contributions: DA, LD, and AD conceived and designed the study. DA, SO, LD, and AD generated and analyzed the data. DA, LD, and AD wrote the manuscript. All authors read and approved the final manuscript.

Acknowledgements: We are grateful to Z.M. Muñoz for discussions during early stages of this study, to A. Fernández-Cortés for help with computational analyses, and to A. Espinosa-Cantú, T. Hernández-Hernández, A. Hernández-López, and L. Riego for critical reading of the manuscript.

References

1. Dehal P, Boore JL: **Two rounds of whole genome duplication in the ancestral vertebrate.** *PLoS Biology* 2005, **3**(10):e314.
2. Adams KL, Wendel JF: **Polyploidy and genome evolution in plants.** *Current Opinion in Plant Biology* 2005, **8**(2):135-141.
3. Wolfe KH, Shields DC: **Molecular evidence for an ancient duplication of the entire yeast genome.** *Nature* 1997, **387**(6634):708-712.
4. Kellis M, Birren BW, Lander ES: **Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*.** *Nature* 2004, **428**(6983):617-624.
5. Sémon M, Wolfe KH: **Consequences of genome duplication.** *Current Opinion in Genetics & Development* 2007, **17**(6):505-512.
6. Papp B, Pal C, Hurst LD: **Dosage sensitivity and the evolution of gene families in yeast.** *Nature* 2003, **424**(6945):194-197.
7. Veitia RA, Bottani S, Birchler JA: **Cellular reactions to gene dosage imbalance: genomic, transcriptomic and proteomic effects.** *Trends in Genetics* 2008, **24**(8):390-397.
8. Seoighe C, Gehring C: **Genome duplication led to highly selective expansion of the *Arabidopsis thaliana* proteome.** *Trends in Genetics* 2004, **20**(10):461-464.
9. Airoidi CA, Davies B: **Gene duplication and the evolution of plant MADS-box transcription factors.** *Journal of Genetics and Genomics* 2012, **39**(4):157-165.
10. Holland PW, Garcia-Fernández J: **Hoxgenes and chordate evolution.** *Developmental Biology* 1996, **173**(2):382-395.
11. Canestro C, Albalat R, Irimia M, Garcia-Fernandez J: **Impact of gene gains, losses and duplication modes on the origin and diversification of vertebrates.** *Seminars in Cell & Developmental Biology* 2013, **24**(2):83-94.
12. Marcet-Houben M, Gabaldón T: **Beyond the whole-genome duplication: Phylogenetic evidence for an ancient interspecies hybridization in the baker's yeast lineage.** *PLoS Biology* 2015, **13**(8):e1002220.
13. Wolfe KH: **Origin of the yeast whole-genome duplication.** *PLoS Biology* 2015, **13**(8):e1002221.
14. Byrne KP, Wolfe KH: **The Yeast Gene Order Browser: combining curated homology and syntenic context reveals gene fate in polyploid species.** *Genome Research* 2005, **15**(10):1456-1461.
15. Conant GC, Wolfe KH: **Increased glycolytic flux as an outcome of whole-genome duplication in yeast.** *Molecular Systems Biology* 2007, **3**:129.
16. van Hoek MJ, Hogeweg P: **Metabolic adaptation after whole genome duplication.** *Molecular Biology and Evolution* 2009, **26**(11):2441-2453.
17. Seoighe C, Wolfe KH: **Yeast genome evolution in the post-genome era.** *Current Opinion in Microbiology* 1999, **2**(5):548-554.
18. Conant GC, Wolfe KH: **Turning a hobby into a job: how duplicated genes find new functions.** *Nature Reviews Genetics* 2008, **9**(12):938-950.
19. Gao L-Z, Innan H: **Very low dene duplication rate in the yeast genome.** *Science (New York, NY)* 2004, **306**(5700):1367-1370.
20. Evangelisti AM, Conant GC: **Nonrandom survival of gene conversions among yeast ribosomal proteins duplicated through genome doubling.** *Genome Biology and Evolution* 2010, **2**:826-834.

21. Scienski K, Fay JC, Conant GC: **Patterns of gene conversion in duplicated yeast histones suggest strong selection on a coadapted macromolecular complex.** *Genome Biology and Evolution* 2015, **7**(12):3249-3258.
22. DeLuna A, Vetsigian K, Shoresh N, Hegreness M, Colon-Gonzalez M, Chao S, Kishony R: **Exposing the fitness contribution of duplicated genes.** *Nature Genetics* 2008, **40**(5):676-681.
23. Dean EJ, Davis JC, Davis RW, Petrov DA: **Pervasive and persistent redundancy among duplicated genes in yeast.** *PLoS Genetics* 2008, **4**(7):e1000113.
24. Byrne KP, Wolfe KH: **Consistent patterns of rate asymmetry and gene loss indicate widespread neofunctionalization of yeast genes after whole-genome duplication.** *Genetics* 2007, **175**(3):1341-1350.
25. Lynch M, Conery JS: **The evolutionary fate and consequences of duplicate genes.** *Science (New York, NY)* 2000, **290**(5494):1151-1155.
26. Kondrashov FA, Rogozin IB, Wolf YI, Koonin EV: **Selection in the evolution of gene duplications.** *Genome Biology* 2002, **3**(2):research0008.0001-research0008.0009.
27. Davis JC, Petrov DA: **Preferential duplication of conserved proteins in eukaryotic genomes.** *PLoS Biology* 2004, **2**(3):E55.
28. Scannell DR, Wolfe KH: **A burst of protein sequence evolution and a prolonged period of asymmetric evolution follow gene duplication in yeast.** *Genome Research* 2008, **18**(1):137-147.
29. Pegueroles C, Laurie S, Albà MM: **Accelerated evolution after gene duplication: A time-dependent process affecting just one copy.** *Molecular Biology and Evolution* 2013, **30**(8):1830-1842.
30. Conant GC, Wagner A: **Asymmetric sequence divergence of duplicate genes.** *Genome Research* 2003, **13**(9):2052-2058.
31. Zhang P, Gu Z, Li WH: **Different evolutionary patterns between young duplicate genes in the human genome.** *Genome Biology* 2003, **4**(9):R56.
32. Brunet FG, Roest Crolius H, Paris M, Aury JM, Gibert P, Jaillon O, Laudet V, Robinson-Rechavi M: **Gene loss and evolutionary rates following whole-genome duplication in teleost fishes.** *Molecular Biology and Evolution* 2006, **23**(9):1808-1816.
33. Ohno S: **Evolution by gene duplication:** Springer Science & Business Media; 1970.
34. Hughes AL: **The evolution of functionally novel proteins after gene duplication.** *Proceedings Biological Sciences / The Royal Society* 1994, **256**(1346):119-124.
35. Wapinski I, Pfeffer A, Friedman N, Regev A: **Natural history and evolutionary principles of gene duplication in fungi.** *Nature* 2007, **449**(7158):54-61.
36. Pál C, Papp B, Hurst LD: **Highly expressed genes in yeast evolve slowly.** *Genetics* 2001, **158**(2):927-931.
37. Drummond DA, Bloom JD, Adami C, Wilke CO, Arnold FH: **Why highly expressed proteins evolve slowly.** *Proceedings of the National Academy of Sciences of the United States of America* 2005, **102**(40):14338-14343.
38. Zhang J, Yang JR: **Determinants of the rate of protein sequence evolution.** *Nature Reviews Genetics* 2015, **16**(7):409-420.
39. Breker M, Gymrek M, Schuldiner M: **A novel single-cell screening platform reveals proteome plasticity during yeast stress responses.** *The Journal of Cell Biology* 2013, **200**(6):839-850.
40. Dyson HJ, Wright PE: **Intrinsically unstructured proteins and their functions.** *Nature Reviews Molecular Cell Biology* 2005, **6**(3):197-208.

41. Brown CJ, Takayama S, Campen AM, Vise P, Marshall TW, Oldfield CJ, Williams CJ, Dunker AK: **Evolutionary rate heterogeneity in proteins with long disordered regions.** *Journal of Molecular Evolution* 2002, **55**(1):104-110.
42. Hirsh AE, Fraser HB: **Protein dispensability and rate of evolution.** *Nature* 2001, **411**(6841):1046-1049.
43. He X, Zhang J: **Rapid subfunctionalization accompanied by prolonged and substantial neofunctionalization in duplicate gene evolution.** *Genetics* 2005, **169**(2):1157-1164.
44. Corrochano LM, Kuo A, Marcet-Houben M, Polaino S, Salamov A, Villalobos-Escobedo JM, Grimwood J, Alvarez MI, Avalos J, Bauer D *et al*: **Expansion of signal transduction pathways in fungi by extensive genome duplication.** *Current Biology* 2016, **26**(12):1577-1584.
45. Gonzalez-Novo A, Labrador L, Pablo-Hernando ME, Correa-Bordes J, Sanchez M, Jimenez J, Vazquez de Aldana CR: **Dbf2 is essential for cytokinesis and correct mitotic spindle formation in *Candida albicans*.** *Molecular Microbiology* 2009, **72**(6):1364-1378.
46. Toyn JH, Johnston LH: **The Dbf2 and Dbf20 protein kinases of budding yeast are activated after the metaphase to anaphase cell cycle transition.** *The EMBO Journal* 1994, **13**(5):1103-1113.
47. Meitinger F, Palani S, Pereira G: **The power of MEN in cytokinesis.** *Cell Cycle (Georgetown, Tex)* 2012, **11**(2):219-228.
48. Treck T, Larson DR, Moldon A, Query CC, Singer RH: **Single-molecule mRNA decay measurements reveal promoter- regulated mRNA stability in yeast.** *Cell* 2011, **147**(7):1484-1497.
49. Conant GC, Wolfe KH: **Functional partitioning of yeast co-expression networks after genome duplication.** *PLoS Biology* 2006, **4**(4):e109.
50. Sprague GF Jr: **Isolation and characterization of a *Saccharomyces cerevisiae* mutant deficient in pyruvate kinase activity.** *Journal of Bacteriology* 1977, **130**(1):232-241.
51. Boles E, Schulte F, Miosga T, Freidel K, Schlüter E, Zimmermann FK, Hollenberg CP, Heinisch JJ: **Characterization of a glucose-repressed pyruvate kinase (Pyk2p) in *Saccharomyces cerevisiae* that is catalytically insensitive to fructose-1,6-bisphosphate.** *Journal of Bacteriology* 1997, **179**(9):2987-2993.
52. Sinha A, Maitra PK: **Induction of specific enzymes of the oxidative pentose phosphate pathway by glucono-delta-lactone in *Saccharomyces cerevisiae*.** *Journal of General Microbiology* 1992, **138**(9):1865-1873.
53. Voordeckers K, Brown CA, Vanneste K, van der Zande E, Voet A, Maere S, Verstrepen KJ: **Reconstruction of ancestral metabolic enzymes reveals molecular mechanisms underlying evolutionary innovation through gene duplication.** *PLoS Biology* 2012, **10**(12):e1001446.
54. Edgar RC: **MUSCLE: multiple sequence alignment with high accuracy and high throughput.** *Nucleic Acids Research* 2004, **32**(5):1792-1797.
55. Suyama M, Torrents D, Bork P: **PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments.** *Nucleic Acids Research* 2006, **34**(suppl 2):W609-W612.
56. Huerta-Cepas J, Serra F, Bork P: **ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data.** *Molecular Biology and Evolution* 2016, **33**(6):1635-1638.
57. Wapinski I, Pfeffer A, Friedman N, Regev A: **Automatic genome-wide reconstruction of phylogenetic gene trees.** *Bioinformatics (Oxford, England)* 2007, **23**(13):i549-558.
58. Yang Z: **PAML 4: phylogenetic analysis by maximum likelihood.** *Molecular Biology and Evolution* 2007, **24**(8):1586-1591.

59. Breslow DK, Cameron DM, Collins SR, Schuldiner M, Stewart-Ornstein J, Newman HW, Braun S, Madhani HD, Krogan NJ, Weissman JS: **A comprehensive strategy enabling high-resolution functional analysis of the yeast genome.** *Nature Methods* 2008, **5**(8):711-718.
60. Landry CR, Levy ED, Michnick SW: **Weak functional constraints on phosphoproteomes.** *Trends in Genetics* 2009, **25**(5):193-197.
61. Jones DT, Ward JJ: **Prediction of disordered regions in proteins from position specific score matrices.** *Proteins* 2003, **53** Suppl 6:573-578.

Figures

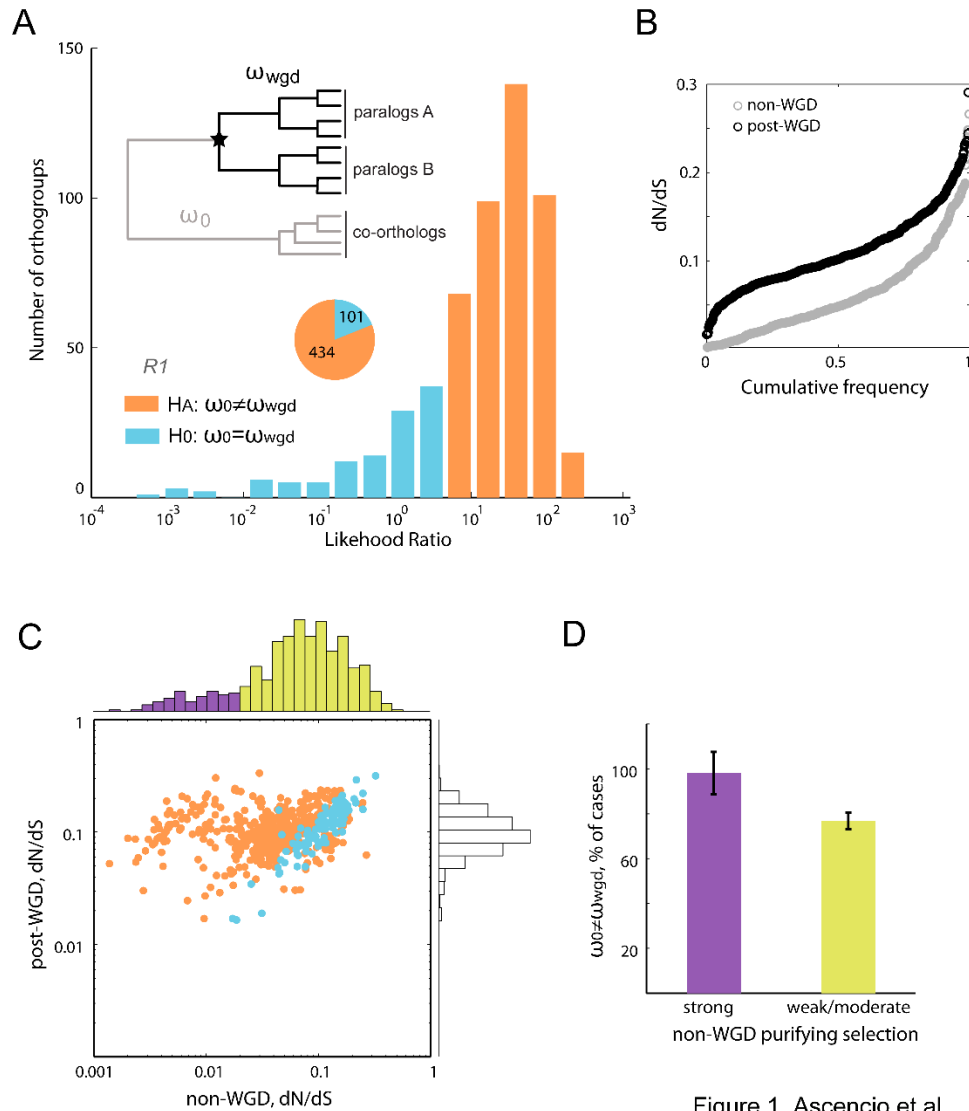


Figure 1_Ascencio et al.

Figure 1. Proteins under strong purifying selection increase their rates of evolution after the WGD. (A) Frequency distribution of the Likelihood-Ratio Test (LRT) values for the hypothesis test comparing two evolutionary models. Orange bars indicate orthogroups in which the null hypothesis of a single dN/dS across the phylogeny is rejected in *R1* (rate-change, $\omega_0 \neq \omega_{wgd}$); blue bars are orthogroups for which the null hypothesis cannot be rejected

(constant-rate, $\omega_0 = \omega_{\text{WGD}}$). **(B)** Cumulative distribution plots comparing the average dN/dS ratios of non- and post-WGD branches. **(C)** Scatter plot comparing post-duplication to pre-duplication dN/dS ratios in 535 orthogroups. Purple bars of the *x*-axis histogram indicate orthogroups under strong purifying selection before the WGD ($\text{dN/dS} < 0.02$) whereas lime bars are those under weak to moderate purifying selection. Each dot is a single orthogroup; orange dots are those in which the null hypothesis is rejected (rate-change, $\omega_0 \neq \omega_{\text{WGD}}$), while cyan dots are orthogroups in which the null hypothesis cannot be rejected. **(D)** Fraction of orthogroups under strong (purple bar) or weak to moderate (lime bar) purifying selection in which the alternative hypothesis (rate-change, $\omega_0 \neq \omega_{\text{WGD}}$) is accepted.

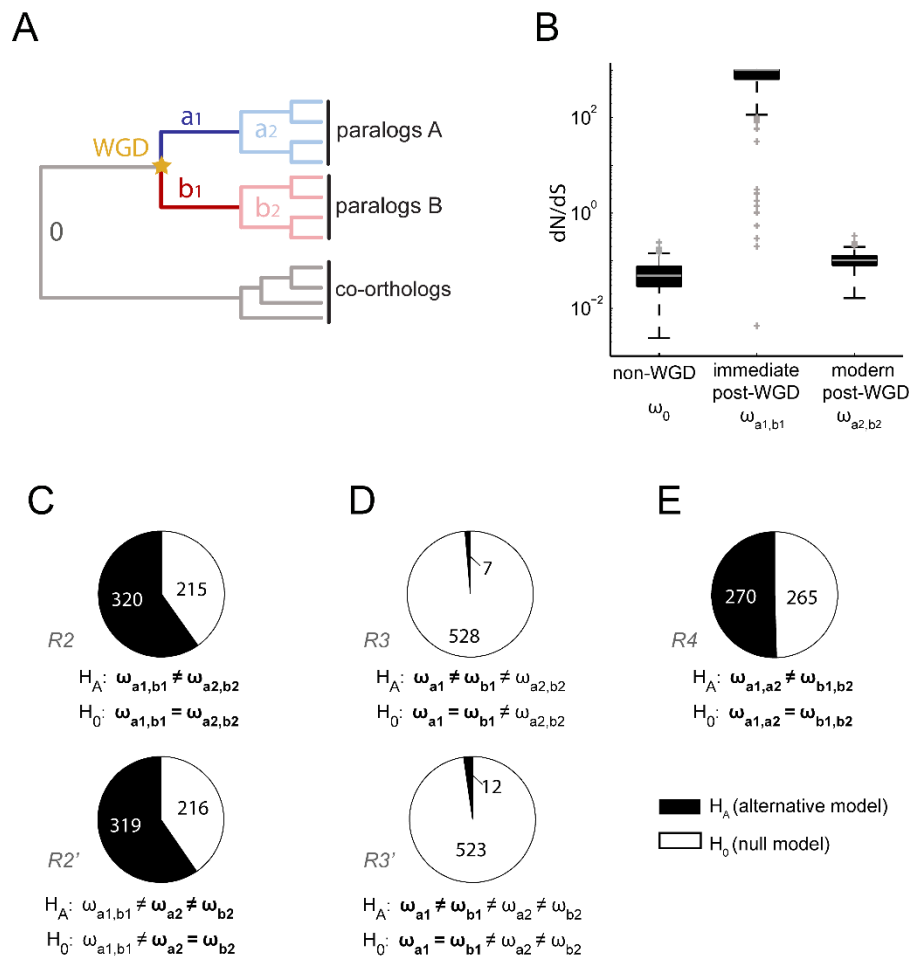


Figure 2. Initial post-WGD acceleration is followed by asymmetric deceleration in protein evolution rates. **(A)** Schematic representing the different branches used in the hypothesis testing for differences in the rates of protein evolution at different branches in the genes' phylogeny. Branches 0 (gray) indicate sequences from all species that diverged before the WGD (yellow star); branches a_1 (blue) and b_1 (red) are the two branches following immediately after the WGD, while a_2 (light blue) and b_2 (pink) are the modern post-WGD paralogous clades. **(B)** Box plots showing the distributions of dN/dS ratios at different stages

of the phylogeny of orthogroups for which the $R2$ hypothesis ($\omega_0 \neq \omega_{a1-b1} \neq \omega_{a2-b2}$) was accepted. Pie charts show the number of orthogroups in which the alternative hypothesis of different evolutionary rates is accepted for (C) immediate post-WGD and modern post-WGD branches ($R2$ and $R2'$ variant), (D) immediate post-WGD paralogous branches ($R3$ and $R3'$ variant), and (E) entire post-WGD paralogous clades ($R4$).

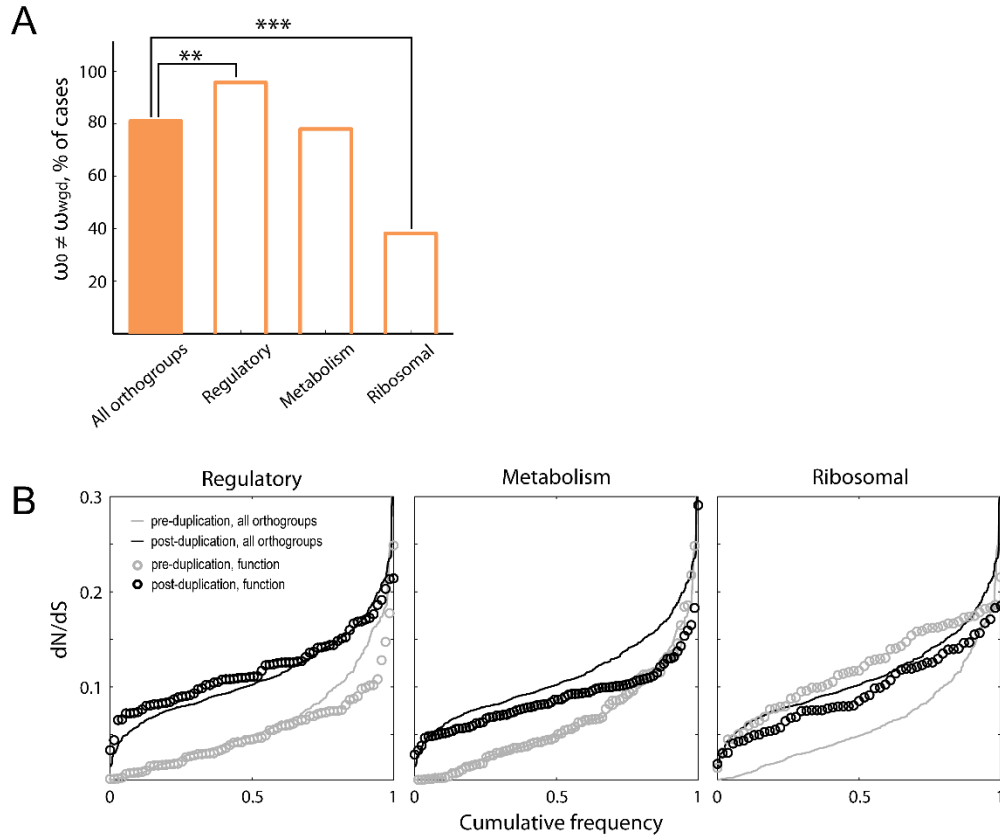


Figure 3. Different protein functions show different patterns of evolution before and after WGD. (A) Fraction of orthogroups for which the alternative hypothesis ($\omega_0 \neq \omega_{\text{wgd}}$) is accepted; plot shows data for all orthogroups and different functional categories. Functional categories which are enriched or depleted in ω -change orthogroups are shown (** $p < 0.01$ and *** $p < 0.001$, χ^2 test). **(B)** Cumulative distribution plots of the dN/dS ratios for three different functional categories. Gray lines (all genes) and circles (specific function) show the pre-duplication dN/dS ratios; black lines and circles indicate post-duplication ratios.

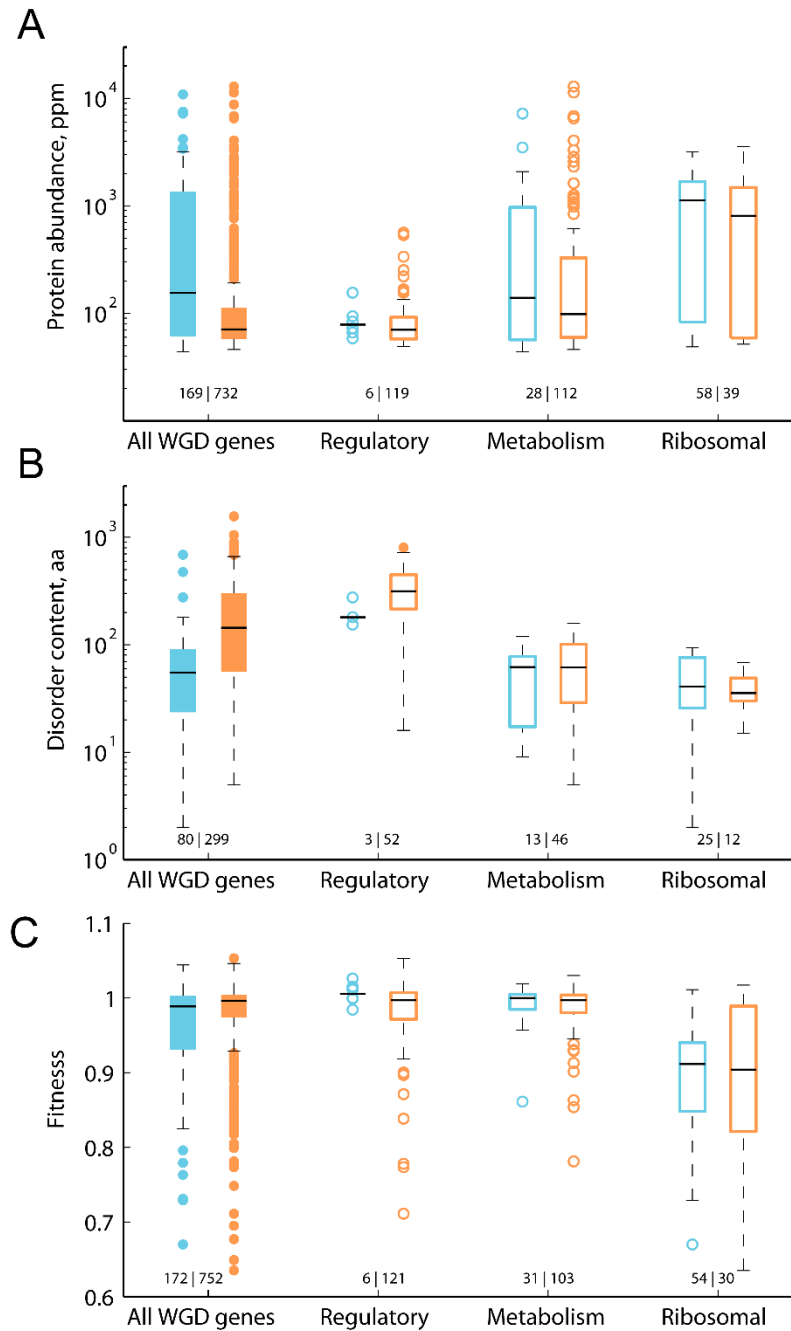


Figure 4. Protein function influences the association of abundance, disordered content, and contribution to fitness with the change in rate of evolution after the WGD. Box plots show *S. cerevisiae* data of (A) absolute protein abundance (ppm), (B) intrinsic disorder content (number of amino acids), and (C) relative fitness of the single-gene deletion for all orthogroups and genes with specific functional annotations. Blue and orange box plots indicate ω -constant and ω -change ($\omega_0 \neq \omega_{\text{wgd}}$) orthogroups, respectively. Numbers below each box plot indicate the corresponding sample size.

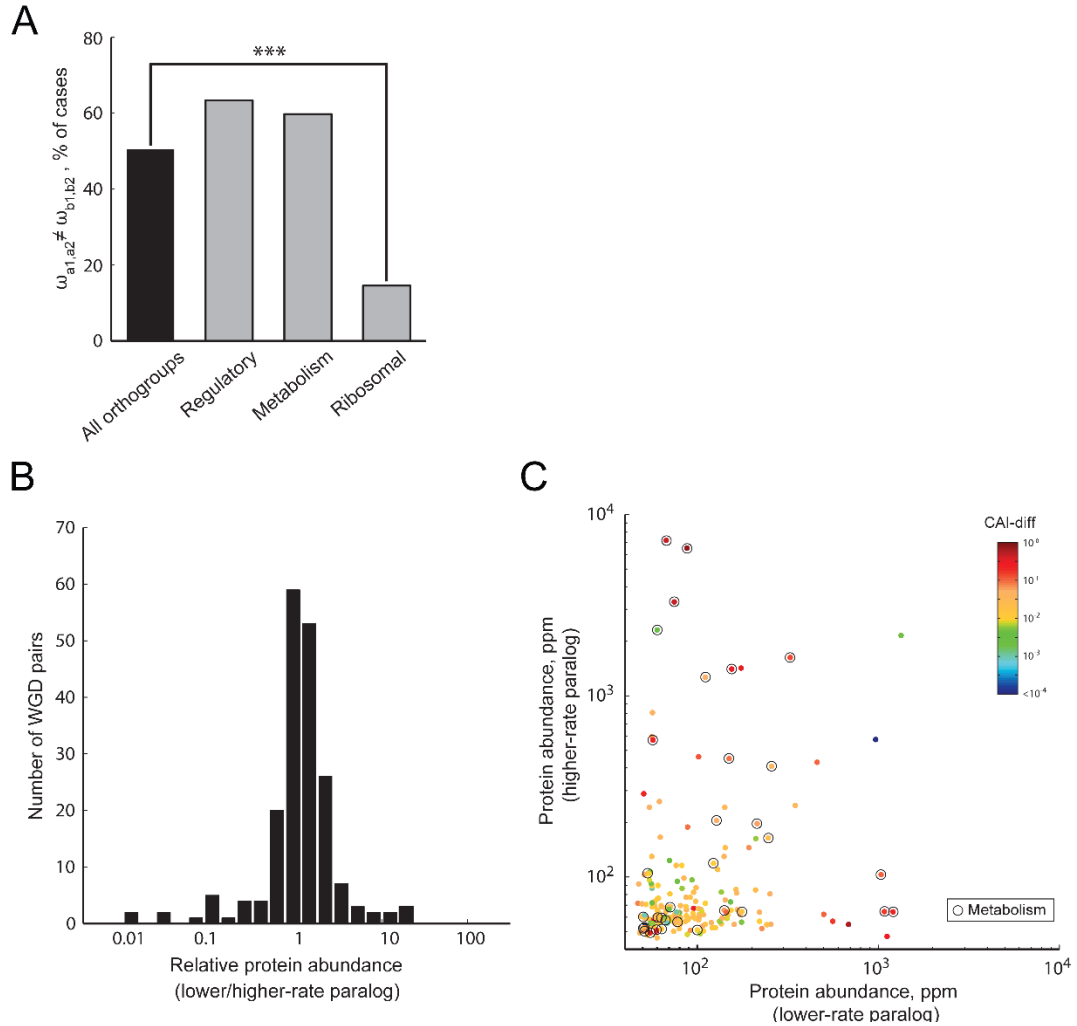


Figure 5. Functional analysis of asymmetric changes in the rates of evolution after WGD. (A) Fraction of orthogroups for which the alternative $R4$ hypothesis ($\omega_{a1,a2} \neq \omega_{b1,b2}$) is accepted; plot shows data for all orthogroups (black bar) and different functional annotations (gray bars). Functional annotations depleted in asymmetric-rate orthogroups are shown ($***p < 0.001$, χ^2 test). (B) Histogram shows the distribution of protein abundances of the low-rate paralog relative to its high-rate paralog. (C) Scatter plot shows absolute protein abundances of the high-rate paralog compared to its low-rate paralog. Color scale indicates the absolute difference between the Codon Adaptation Indexes of the paralogs (CAI-diff). Black circles indicate all genes encoding for metabolic enzymes.

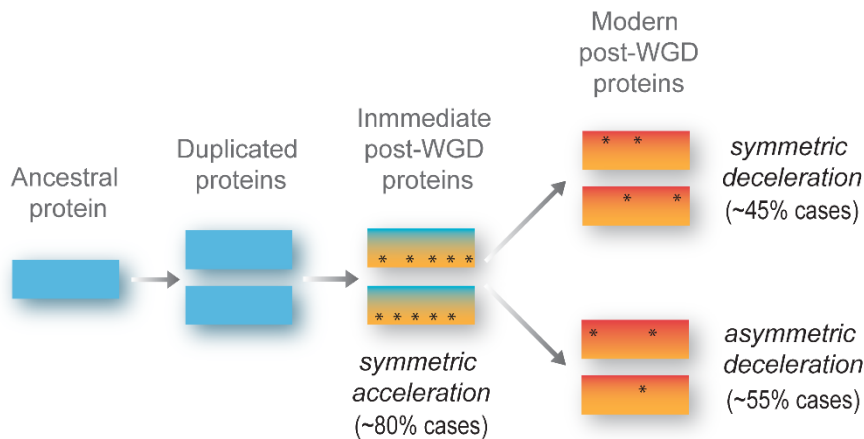


Figure 6. Stages of protein evolution after WGD. Immediately after duplication both paralogous copies undergo an initial phase of accelerated rate of evolution that affects the two gene copies in a similar manner (symmetric loss of evolutionary constraint), followed by a phase of deceleration in the rates of evolution (return to constraint). The degree of deceleration can be similar for both copies (symmetric return to constraint) or different between the paralogous copies (asymmetric return to constraint). Asterisks illustrate non-synonymous substitutions in the genes' sequences.

Additional files

Table S1. YGOB data set analysis, including dN/dS data, likelihood-ratio tests' results, and *S. cerevisiae* protein properties and functions for 535 orthogroups.

Table S2. FOR data set analysis, including dN/dS data, likelihood-ratio tests' results for 462 orthogroups.

Figure S1. Analysis of evolution rates in fungal WGD orthogroups. The *S. cerevisiae* *PSK1/PSK2* orthogroup is shown as an example of significantly relaxed purifying selection after duplication. Gray branches indicate the orthologs of *PSK1/PSK2* in ascomycetes species that diverged before the WGD (yellow star), while black branches indicate post-WGD species with two paralogous clades. Average dN/dS ratios are shown for both non-WGD (ω_0) and post-WGD genes (ω_{wgd}); the *p*-value is from the likelihood ratio test of the alternative hypothesis of different rates of protein evolution between ω_0 and ω_{wgd} .

Figure S2. Results are consistent for most orthogroups when using two alternative data sets. (A) Comparison of Likelihood-Ratio Test (LRT) values for the hypothesis test comparing the two evolutionary models (null: $\omega_0 = \omega_{\text{wgd}}$; alternative: $\omega_0 \neq \omega_{\text{wgd}}$) using sequences from two different datasets: the Yeast Genome Order Browser (YGOB) [14] and the Fungal Orthogroups Repository (FOR) [35]. (B) Venn diagram of orthogroups in which the alternative hypothesis ($\omega_0 \neq \omega_{\text{wgd}}$) is accepted in the YGOB and the FOR data sets, (C) Comparison of non-WGD (gray dots) and post-WGD dN/dS ratios of 426 orthogroups calculated for sequences for the YGOB and the FOR data sets.

ANEXO G

Artículo 2: en preparación

Coupled Divergence of Coding and *Cis*-Regulatory Functions after Gene Duplication

Diana Ascencio & Alexander DeLuna

Laboratorio Nacional de Genómica para la Biodiversidad (LANGEBIO), Centro de Investigación y de Estudios Avanzados (CINVESTAV), 36821 Irapuato, Mexico

Correspondence to:

Alexander de Luna

Laboratorio Nacional de Genómica para la Biodiversidad

Centro de Investigación y de Estudios Avanzados del IPN

Km 9.6 Libramiento Norte Carretera Irapuato-León

36821 Irapuato, Guanajuato, Mexico

Tel: +52 462 166-3009

E-mail: adeluna@langebio.cinvestav.mx

ABSTRACT

Changes in coding (protein) and in *cis*-regulatory (promoter) regions underlie evolution after gene duplication. However, it is unclear what is the relative functional contribution of each genetic region to adaptive evolution, and if such divergence of proteins and promoters is coupled or independent. Here, we use chimeric constructs of coding and *cis*-regulatory sequences to systematically measure the fitness contribution of diverged duplicated proteins and promoters. We find that the degrees of phenotypic compensation between the two regions are correlated, suggesting coupled divergence of coding and *cis*-regulatory functions after gene duplication. Surprisingly, paralogous proteins typically provide a lower degree of functional compensation than their corresponding paralogous promoters. While gene expression diverges rapidly after duplication, our experimental results suggest that changes in coding sequences usually underlie the long-term evolution by gene duplication in yeast.

Following gene duplication, paralogs diverge in their coding (protein) and *cis*-regulatory (promoter) sequences. Such divergence leads to changes in gene function and, ultimately, to the evolution of phenotypic traits {Ohno, 2013 #58; Hughes, 1994 #60; Zhang, 2003 #62}. While several inferences of the evolutionary roles of protein and promoter divergence have been suggested {Castillo-Davis, 2004 #11; Stern, 2008 #15; Carroll, 2005 #10}, the relative adaptive value of these genetic regions has not been measured directly. Moreover, the correlation between divergence in protein and promoter functions remains unclear. Diverged duplicate genes may fall into one of three distinct groups: (1) Paralogs with diverged promoter function and conserved protein function, (2) diverged protein function and conserved promoter function, or (3) coupled diverged protein and promoter functions (**fig. 1A**).

To measure the contribution of proteins and promoters to phenotypic diversification after gene duplication, we developed a strategy based on the competitive fitness profiling of genetically-engineered *Saccharomyces cerevisiae* strains (**fig. 1B**). By complementing knock-out strains with knock-in constructs of paralogous sequences, we tell apart the individual contributions of proteins and promoters to the non-overlapping (diverged) function. From such high-resolution fitness data, we define a protein functional divergence index (D_{orf}) and a corresponding index of promoter divergence (D_{prm}) for each

paralogous pair (**fig. 1C-D**; see Materials and Methods). Such functional-divergence indexes provide direct and comparable metrics to quantify the extents to which coding and promoter sequences have contributed to phenotypic novelty after gene duplication.

To identify duplicate gene pairs with non-overlapping functions, we screened a collection of 711 single-gene knockouts of non-essential paralogs originated from the whole-genome duplication (WGD). Duplicates from the WGD provide an unambiguous model in which XXX and both XXXX. We identified 31 *X1/2* duplicate pairs in which one of the genes (herein *X1*) was indispensable—i.e. knockout of the gene was strongly deleterious—while its paralog (*X2*) was dispensable—i.e. deletion resulted in no significant fitness loss (~10% of pairs tested; see **supplementary fig. S1, supplementary table S1**). Such asymmetric phenotypic profiles are indicative of loss of functional overlap whereby different biochemical or *cis*-regulatory properties render the dispensable genes unable to compensate for the lack of their indispensable paralogs under a specific experimental condition.

We focused our initial comparative functional-divergence analysis on seven duplicate-gene pairs with asymmetric deletion phenotypes: *BAG7/SAC7*, *CLA4/SKM1*, *GUP1/2*, *TIF4631/2*, *TSA1/2*, *SET3/4*, and *VPS5/YKR078W*. These genes include duplicate pairs with different degrees of coding-sequence

conservation (20–85% amino acid identity) and varied functional annotations (**supplementary table S2**). For each gene pair analyzed, we generated paralogous ORF, promoter (*5'UTR*), and control constructs (**supplementary fig. S2**, see Materials and Methods). To estimate D_{orf} and D_{prm} under different functional contexts, we measured fitness of all constructs under SC medium and three environmental-stress conditions (**supplementary fig. S3**). Our competitive-based fitness profiling method showed good technical replicability (**supplementary fig. S4**). As expected, knockout mutations at both ORF and promoter regions typically had similar effects on fitness, while the *XI* knock-ins of ORFs or promoters usually complemented the slow-growth phenotypes close to WT growth (**supplementary fig. S5**).

The fitness profiles of the ΔxI knock-out and knock-in strains (*XI* controls and *X2* and paralogs for both coding and promoter sequences) are shown in **fig. 2A**. We observed that paralogous *X2* sequences typically provided a lower degree of phenotypic complementation than the *XI* knock-in controls, which reflects some degree of functional divergence between paralogs. Nonetheless, several instances resulted in full complementation by the paralogous gene region, suggesting little divergence at the functional level. From the measured complementation patterns, we calculated D_{orf} and D_{prm} for all gene pairs under each environmental condition tested. When comparing the

complementation capacity driven by *X2* protein and promoter sequences, we noted that in many cases such sequences provided similar degrees of complementation. Interestingly, specific gene pairs/environments showed highly dissimilar patterns depending on whether the paralogous coding or promoter sequence was used, suggesting a functional-divergence bias in such regions. It was more common to observe that the paralogous promoter expressing the endogenous protein resulted in higher fitness than the paralogous protein driven by the endogenous promoter. Noteworthy, almost all instances of strong promoter divergence were coupled to strong functional divergence at the protein level.

The degrees of functional divergence of paralogous coding and *cis*-regulatory sequences, inferred by D_{orf} and D_{prm} of each gene pair under each environmental condition, showed to be significantly correlated (**fig. 2B**, $p=0.006$). We also computed divergence indexes for each of the seven gene pairs using an integrated fitness value across all conditions tested. This suggests that functional changes in proteins and promoters contribute to fitness in a coupled manner, namely XXXXX.

Importantly, several data points fell below the $D_{orf} = D_{prm}$ diagonal, suggesting that protein sequences tend to have a larger contribution to functional divergence than the corresponding promoter sequences. Finally, we note that high degrees of functional divergence of paralogous

protein sequences were not necessarily associated to fast evolution at the protein-sequence level (see Ka values in **fig. 2B**).

The above observations hold for a set of duplicate genes which have diverged in such a way that maximum fitness cannot be reached when the *X1* paralog is missing. However, such knockouts are not lethal, either because the dispensable *X2* paralog can provide partial compensation for the lack of *X1* and hence support cell growth, or because the function of *X1* is not completely essential for growth. Since essential paralogs provide unambiguous scenarios of complete phenotypic divergence, we evaluated the functional compensation capacity provided by paralogous proteins or promoters of essential WGD duplicate genes. To this end, we assayed viability of meiotic derivatives bearing knock-out or knock-in constructs at seven such duplicate genes (**fig. 2C**; **supplementary fig. S6**). We observed that in four out of seven cases, none of the individual genetic regions could revert the lethal phenotype of the $\Delta x1$ knockout, suggesting high functional divergence of both proteins and promoters. These results further suggest that phenotypic evolution of duplicate-gene functions is typically driven by coupled divergence in their coding and *cis*-regulatory regions.

Even though it is well documented that *cis*-regulatory sequences diverge rapidly after duplication {Gu, 2002 #33; Li, 2005 #27; Tirosh, 2007 #25}, our functional assays indicated that many paralogous

promoters are readily exchangeable, even after long evolutionary times. In a trivial scenario, the high compensation capacity of paralogous promoters could reflect low requirements of gene expression, or that any given level of expression of the *X1* protein is enough to achieve WT fitness. To directly measure the relationship between expression level and fitness, we generated *tet₀₇-promoter* (*tet_{prm}*) replacements of the indispensable *X1* paralogs (**fig. 3A-G**). In all cases but one (*TSA1*), we found that fitness depended strongly on gene-expression level, and that the synthetic promoter was unable to provide full functional compensation at a wide range of expression levels of the indispensable *X1* paralog. As expected, changes in the expression of the *X2* paralogs generally resulted in no fitness cost—*SET4* was an interesting exception in which high expression was associated to fitness below the WT. Flow-cytometry analysis of a *tet_{prm}*-driven GFP reporter showed that the system allows screening of gene-expression levels of over two orders of magnitude (**fig. 3H**). These results suggest that the observed high compensation capacity is specific of paralogous promoters, which likely reflects partially-overlapping *cis*-regulatory functions.

The quantitative estimate of divergence provided by D_{orf} and D_{prm} indicated several instances of similar levels of loss of functional overlap between the proteins and promoters of several WGD paralog pairs. We found that the *Set4* protein expressed by the *SET3_{prm}* did not restore

fitness of $\Delta set3$; neither did $SET4_{prm}$ while driving expression of the endogenous Set3 protein. *SET3* codes for the defining member of the histone-deacetylase complex Set3C {Pijnappel, 2001 #47}; no role has been defined for *SET4*, which conserves the SET domain but shares little amino acid sequence identity with *SET3*. We have cases of functional divergence of both cis-regulatory and coding region. For instance, both Tsa1p and Tsa2p are peroxiredoxins that cooperate together in the cellular defense against reactive oxygen and nitrogen species {Chae, 1993 #52; Park, 2000 #53; Wong, 2002 #54}. Under normal growth conditions, Tsa1 functions mostly as an antioxidant protecting both the cytoplasm and translational apparatus against endogenous ROS. Over-oxidation of the peroxidatic cysteine residue causes a shift to the high-molecular-mass super-chaperone form. Thus, in response to an acute oxidative stress, a portion of Tsa1 is shifted towards its super-chaperone form which protects against stress denatured proteins {Trotter, 2008 #55}. This is especially important since we found that deletion of *tsa1* is deleterious on oxidative stress condition, and nor the protein or the promoter of *tsa2* were able to complement in such environment. Note high CAI bias, which could explain lack of compensation of seemingly redundant proteins. Likewise, viability of several essential genes was not restored by neither their paralogous proteins nor promoters, indicating functional divergence of both genetic regions. For instance, neither *Dbp1* nor $DBP1_{prm}$ rescued $\Delta ded1$ lethality, an

ATP-dependent DEAD-box RNA helicase {Iost, 1999 #43}, despite considerable amino acid sequence identity (69%) with the paralogous Ded1, surprisingly it has been found that even when the human orthologue *DDX3X* (52% identical) complements the $\Delta ded1$ null mutation {Senissar, 2014 #42}.

Only two paralogous pairs (*GSP1/2* and *PET9/AAC3*) showed full functional divergence exclusively at the promoter level (i.e. the paralogous protein but not the promoter restored growth). This result was expected due to very low rate of non-synonymous substitutions between the paralogs (see *Ka* values in **fig. 2C**). Indeed, it has been reported that growth of $\Delta pet9$ cells is readily restored by overexpression of its *AAC3* paralog {Lawson, 1988 #44}, while the levels of the *GSP2* transcript are about 10-fold lower than those of *GSP1* under glucose {Belhumeur, 1993 #46}. It is worth noting that functional compensation by the redundant protein is achieved despite the high CAI bias in these paralogous pairs. Conversely, the *RSC6/SNF12* paralog pair showed full functional divergence only at the protein level consistent with the high degree of divergence of their amino acid sequences.

While previous studies have estimated the sequence and functional evolutionary changes of proteins and promoters, the fitness changes associated to such molecular divergence remain essentially unknown. Our high-resolution experimental study suggests that phenotypic evolution of duplicate genes is

usually driven by coupled functional divergence in their coding and *cis*-regulatory regions. Despite widespread evidence suggesting that *cis*-regulatory sequences and gene regulation diverge rapidly after gene duplication {Gu, 2002 #33;Li, 2005 #27;Tirosch, 2007 #25}, we find that almost all instances of strong promoter divergence were coupled to strong functional divergence at the protein level. Our results indicate that paralogous promoters in yeast are readily exchangeable, suggesting that the capacity to compensate for each other's loss is maintained, despite rapid sequence divergence. This is in agreement with a scenario of need-based paralogous responsiveness, whereby key regulatory elements shared by paralogous genes grant expression when the function is needed and lower pleiotropy (more robust){DeLuna, 2010 #17}. In fact, our data reflects a bias to stronger fitness contribution associated to changes in coding sequences. Our study highlights the importance of taking into account fitness to gain a better understanding of the molecular traits that drive phenotypic evolution and biological diversification. Our quantitative framework of functional divergence is readily applicable to other organisms or homologous sequences.

MATERIALS AND METHODS

Strains, plasmids and media used. For the phenotypic profiling of the Whole Genome Duplication (WGD) paralogs we

used the collection of single-knockout mutants (with Y7092-YFP genetic background) generated in a previous work {DeLuna, 2008 #3}, and Y7092-CFP as a reference (see Table S3). For the construction of the construction of the chimerical-strains we employed the plasmid pFA6-*kanMX4* {Wach, 1994 #2}, using the *XmaI* restriction site (see Fig. S2). The Y7092-YFP strain was used as parental strain for non-essential paralogous pairs {DeLuna, 2008 #3}. For the duplicate gene pairs with an essential member, the chimeric-strains were constructed as heterozygous diploids (using the YDA100 starter strain). Finally, the construction of strains expressing genes under the *tetO7* system was done with R1158-CFP starter strain.

Selection media for the transformations was Yeast Nitrogen Base without amino acids and ammonium sulfate (BD® Bioscience), with 1g/L of monosodium glutamate as nitrogen source and supplemented the required composition of amino acids and antibiotics. The competition media was low-fluorescence Yeast Nitrogen Base (YNB-lf) modified from a previous work {Sheff, 2004 #4}. For the stress calibration assay, YNB-lf as a base accompanied with NaCl, ethanol or hydrogen peroxide, each condition in three different concentrations (*fig. S3*).

Genetic constructs. We used the starter strain Y7092-YFP from De Luna's work {DeLuna, 2008 #3} to construct all the chimeric strains. First, we cloned all the paralogous X2 genes into the pFA6-

kanMX4 plasmid (see Figure S2, Table S2). Then, we amplified the promoter-sequence or the coding-sequence, from the previously generated plasmids. These sequences contained also the *kanMX4* selection marker, which confers geneticin resistance. After that, we replace the wild-type sequences with the paralogous sequences, generating a set of chimerical strains (**Figure S2, Table S3**). For the positive controls, we used the plasmids containing the *XI* gene sequences. For the knockout strains, we used the empty plasmid (pFA6-*kanMX4*) and DNA template and we only amplified the selection marker. Few other constructs were discarded because one or more strains did not pass a quality-control check (eg. a gene knock-in did not complement wild-type fitness, see Materials and Methods).

We constructed strains containing paralogous genes expressed under the *tet-off* tunable-promoter system {Bellí, 1998 #1}. We used the starting strain R1158-CFP expressing the *tTA* transactivator, necessary *tetO₇* system operation. We used pCM325 {Yen, 2003 #5} as template to amplify the *tetO₇*-cassette (with KanR as a selectable marker), to replace the native promoter(-600bp to -45 bp) via PCR-based directed recombination(see **Table S4**). Positive mutants were selected on YPAD with 200mg / L geneticin (G-418 by Gibco®) and 100mg / L nourseothricin (Werner BioAgents).

SGA-based viability test. For the essential paralogous pairs, we adapted the

synthetic genetic array (SGA) technology {Tong, 2001 #7}. First, we obtained MATa meiotic progeny derived from sporulation of our heterozygous diploids. Once we obtained the haploids we made two consecutive selections first with nourseotricin (linked CFP) and then for both nourseotricin and geneticin (linked with the genetic variant of our interest). The last selection step was made on liquid media in order to evaluate viability by drop spot assay (see **Figure S6**). All colony arrays were transferred manually with a 384-head pin tool (V&P Scientific).

Competitive fitness profiling. The competition fitness assay was adapted from a previous work {DeLuna, 2008 #3}. First, we inoculated 150 µl of fresh competition medium (using Corning® Costar® 96-well cell culture plates) with overnight cultures of mutant and reference strains (tagged with YFP and CFP respectively) in a 1:2 proportion. Every 24hrs, we diluted the saturated competition cultures 16-fold into sterile fresh competition medium. Cultures were propagated in parallel for approximately 28 generations (7 days) in a fully automated robotic system (Freedom EVO, Tecan Ltd) coupled with a microplate multireader (Infinite Reader M100, Tecan Ltd). We monitored every 2hrs the OD600nm, YFP fluorescence signal (515/528nm, gain:130) and CFP fluorescence signal (433/475nm, gain:145). The equipment and cultures were maintained at 30°C and 70% relative humidity. We performed three technical replicas.

Flow cytometry. Flow cytometry was performed as previously described {DeLuna, 2010 #17}. In brief, a flow cytometer with a high-throughput autosampler (LSRII with HTS, Becton Dickinson) was used to record fluorescence from the GFP fluorophore. Cells were measured in high-throughput mode at a flow rate of 0.5 μ l/s for 15 s.

DECLARATIONS

Acknowledgements: We are grateful to XXX for critical reading of the manuscript.

Funding: This work was funded by the Consejo Nacional de Ciencia y Tecnología de México (CONACYT grant CB2011/164889 to AD). DA had a doctoral scholarship from CONACYT (267799).

Ethics approval and consent to participate: Not applicable.

Consent for publication: Not applicable.

Competing interests: The authors declare that they have no competing interests.

Authors' contributions: DA and AD conceived and designed the study. DA did the experiments. DA and AD analyzed the data. DA and AD wrote the manuscript. All authors read and approved the final manuscript.

FIGURES

Fig.1_Ascencio

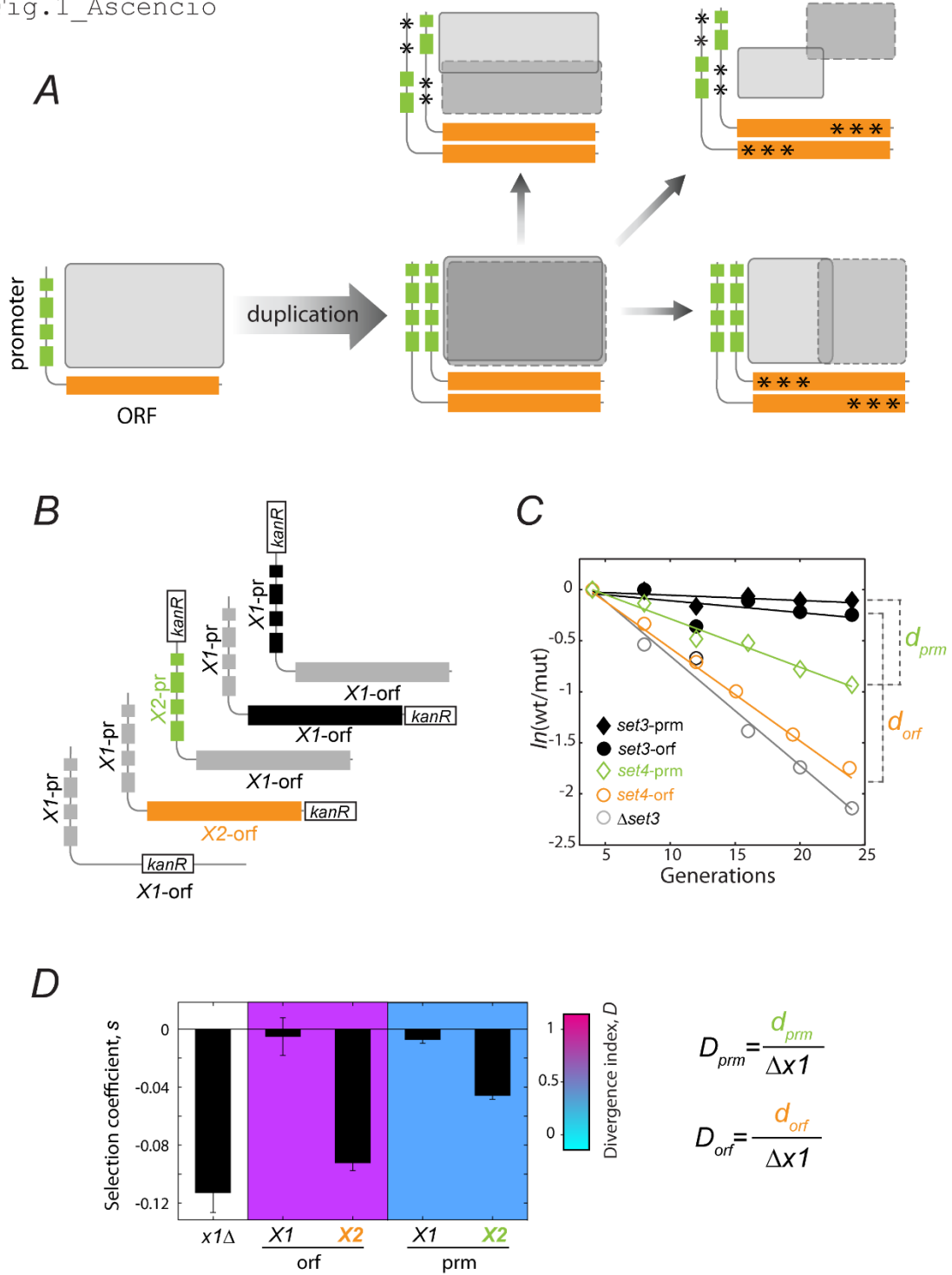


Figure 1. A new phenotypic index for functional diversification of paralogous proteins and promoters. a) The different scenarios of functional divergence of

promoter and protein of duplicate gene pairs. We represent function in two dimensions; regulatory function(promoter) and biochemical function(protein) The rectangles represent such functional space; the height represents the promoter function and the protein function is represented by the width of the rectangle. Immediately after the duplication both genes have full functional overlap. After divergence, the overlap in promoter function, protein function or both is reduced, which leads to the three different scenarios of functional divergence presented here. b) Genetic complementation strategy. On the left, genetic complementation constructs, each complementation was done directly in chromosome on the X1 locus. From top to bottom, the first two constructions are positive controls for promoter and protein replacement, the two in the middle are the chimerical strains of paralogous promoter and protein complementation, the last construction is the X1 loss of function mutant. On the right, representative a fitness complementation assay of SET3/4. We show the ratio between the reference and each complementation strain. The slope difference (selection coefficient) difference between the complementation of promoter or protein and the respective control is used to calculate the divergence index (D_{prm} or D_{orf}) for each region. c) We calculate a divergence index based on the data obtained from the fitness complementation assay. A value close to 0 of the divergence index (D_{prm} or D_{orf}) is indicative of low functional divergence meanwhile a value close to 1 denote high functional divergence.

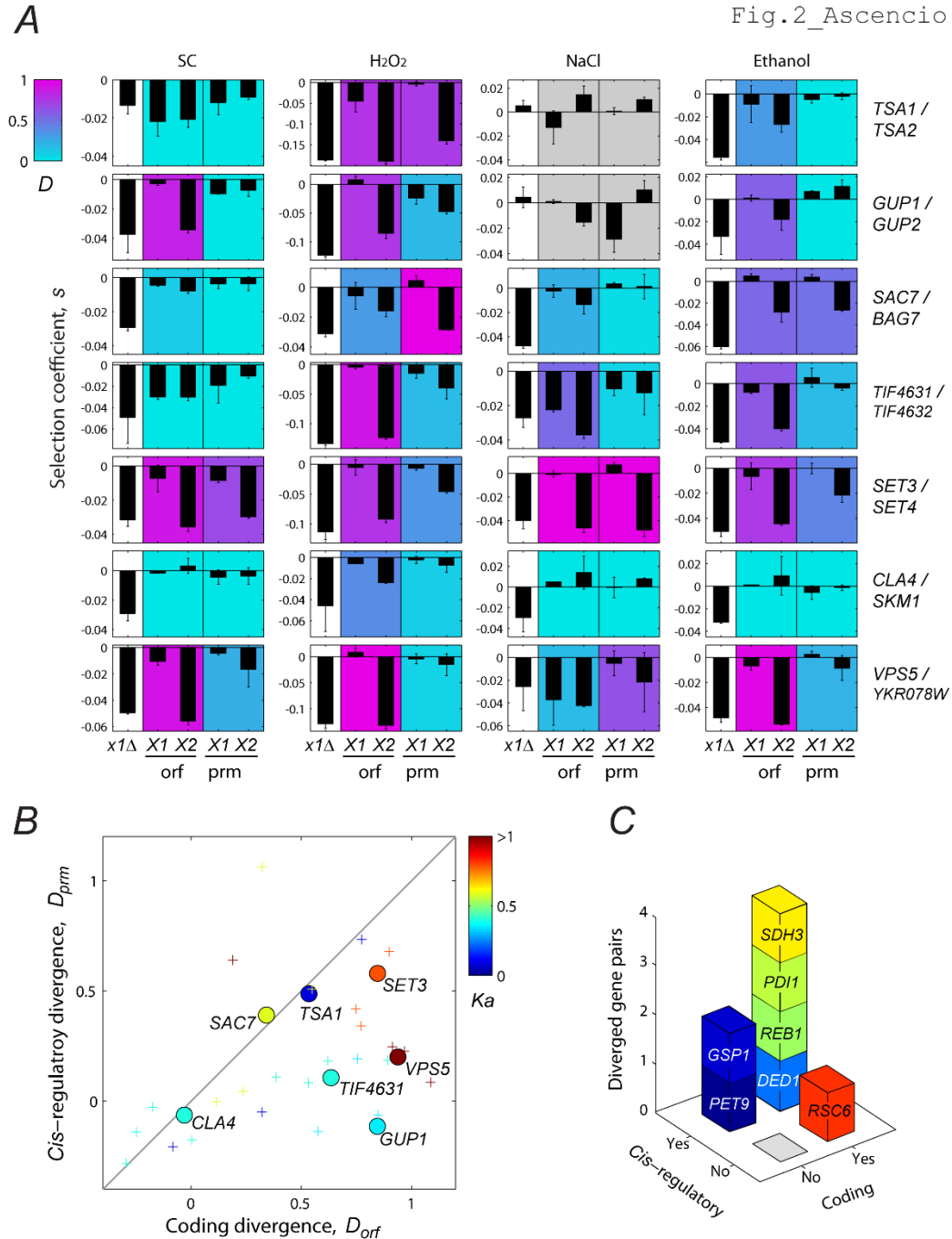


Figure 2. The fitness impact of coding and *cis*-regulatory divergence after gene duplication. a) Competitive fitness profiling of seven paralogous gene-pairs (rows) in four different conditions (columns). In each box we represent five bars, corresponding to the five different genetic constructs, the error bar is represents the standard error of the mean

of three different biological replicas. The divergence index is indicated with a color gradient, from high functional divergence (pink) to lower functional divergence (blue). **b)** Comparison of the promoter and protein divergence index. The divergence indexes for the seven gene pairs in the four conditions are indicated with plus signs, the mean of the four conditions for a given gene-pair is represented with a circle, the color indicates of the non-synonymous substitution rate (Ka) of the coding sequences of the two members of the paralogous gene couple. **c)** Viability complementation assay for seven paralogous pairs with essential members. If the genetic complementation with the paralogous promoter or protein is viable we consider that there is not functional divergence. On the other hand, when the chimerical constructs are not viable is an indicative of high functional divergence. Considering the viability results of the paralogous genetic complementation of both promoter and protein we defined the position of each gene pair.

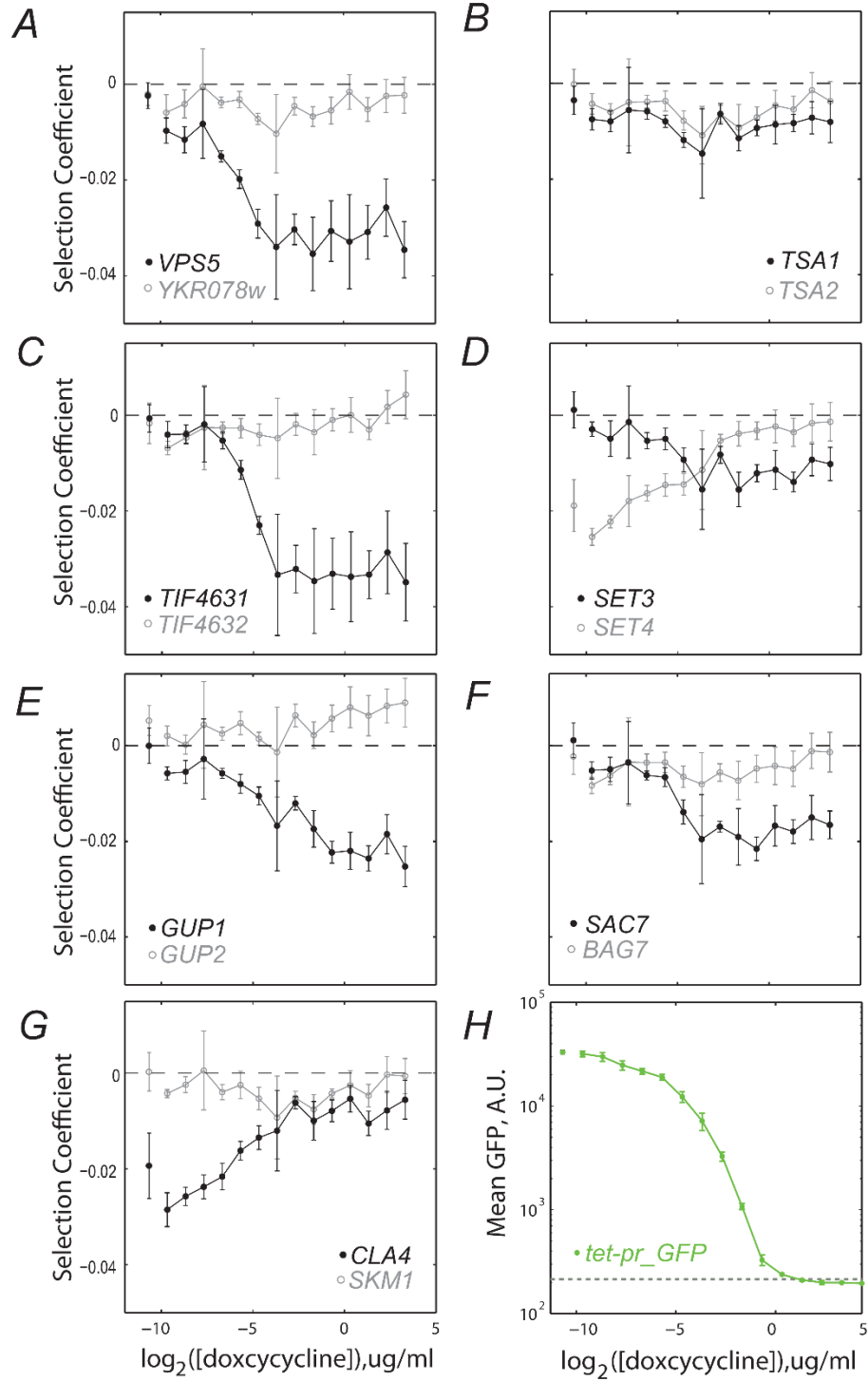


Figure 3. Synthetic promoters provide functional complementation only at narrow levels of gene expression. (a-g) Plots show the selection coefficients of strains with tunable promoters as a function of doxycycline concentration. In each panel, the tetO7

promoter (tet-off system) drives the expression of the indispensable (*X1*, black) or dispensable (*X2*, gray) paralog; its duplicate remains intact. **h**) The *tet-off* system allows over two orders of magnitude control of gene expression. Mean expression level of tetO7-driven YFP reporter at increasing doxycycline concentrations. The dashed line indicates the average signal of the parental strain (auto-fluorescence).

REFERENCES

- Belhumeur P, Lee A, Tam R, DiPaolo T, Fortin N, Clark MW 1993. GSP1 and GSP2, genetic suppressors of the *prp20-1* mutant in *Saccharomyces cerevisiae*: GTP-binding proteins involved in the maintenance of nuclear organization. *Molecular and Cellular Biology* 13: 2152-2161. doi: 10.1128/mcb.13.4.2152
- Bellí G, Garí E, Piedrafita L, Aldea M, Herrero E 1998. An activator/repressor dual system allows tight tetracycline-regulated gene expression in budding yeast. *Nucleic acids research* 26: 942-947.
- Carroll SB 2005. Evolution at two levels: on genes and form. *PLoS Biol* 3: e245. doi: 10.1371/journal.pbio.0030245
- Castillo-Davis CI, Hartl DL, Achaz G 2004. cis-Regulatory and protein evolution in orthologous and duplicate genes. *Genome Res* 14: 1530-1536. doi: 10.1101/gr.2662504
- Chae HZ, Kim IH, Kim K, Rhee SG 1993. Cloning, sequencing, and mutation of thiol-specific antioxidant gene of *Saccharomyces cerevisiae*. *Journal of Biological Chemistry* 268: 16815-16821.
- DeLuna A, Springer M, Kirschner MW, Kishony R 2010. Need-based up-regulation of protein levels in response to deletion of their duplicate genes. *PLoS Biol* 8: e1000347. doi: 10.1371/journal.pbio.1000347
- DeLuna A, Vetsigian K, Shores N, Hegreness M, Colon-Gonzalez M, Chao S, Kishony R 2008. Exposing the fitness contribution of duplicated genes. *Nat Genet* 40: 676-681. doi: 10.1038/ng.123
- Gu Z, Nicolae D, Lu HH, Li WH 2002. Rapid divergence in expression between duplicate genes inferred from microarray data. *Trends Genet* 18: 609-613.
- Hughes AL 1994. The Evolution of Functionally Novel Proteins after Gene Duplication. *Proceedings of the Royal Society of London. Series B: Biological Sciences* 256: 119-124. doi: 10.1098/rspb.1994.0058
- Iost I, Dreyfus M, Linder P 1999. Ded1p, a DEAD-box Protein Required for Translation Initiation in *Saccharomyces cerevisiae*, Is an RNA Helicase. *Journal of Biological Chemistry* 274: 17677-17683. doi: 10.1074/jbc.274.25.17677
- Lawson JE, Douglas MG 1988. Separate genes encode functionally equivalent ADP/ATP carrier proteins in *Saccharomyces cerevisiae*. Isolation and analysis of AAC2. *Journal of Biological Chemistry* 263: 14812-14818.

- Li WH, Yang J, Gu X 2005. Expression divergence between duplicate genes. *Trends Genet* 21: 602-607. doi: 10.1016/j.tig.2005.08.006
- Ohno S. 2013. *Evolution by gene duplication*: Springer Science & Business Media.
- Park SG, Cha M-K, Jeong W, Kim I-H 2000. Distinct Physiological Functions of Thiol Peroxidase Isoenzymes in *Saccharomyces cerevisiae*. *Journal of Biological Chemistry* 275: 5723-5732. doi: 10.1074/jbc.275.8.5723
- Pijnappel WWMP, Schaft D, Roguev A, Shevchenko A, Tekotte H, Wilm M, Rigaut G, Séraphin B, Aasland R, Stewart AF 2001. The *S. cerevisiae* SET3 complex includes two histone deacetylases, Hos2 and Hst1, and is a meiotic-specific repressor of the sporulation gene program. *Genes & Development* 15: 2991-3004. doi: 10.1101/gad.207401
- Senissar M, Saux AL, Belgareh-Touzé N, Adam C, Banroques J, Tanner NK 2014. The DEAD-box helicase Ded1 from yeast is an mRNP cap-associated protein that shuttles between the cytoplasm and nucleus. *Nucleic acids research* 42: 10005-10022. doi: 10.1093/nar/gku584
- Sheff MA, Thorn KS 2004. Optimized cassettes for fluorescent protein tagging in *Saccharomyces cerevisiae*. *Yeast* 21: 661-670. doi: 10.1002/yea.1130
- Stern DL, Orgogozo V 2008. The loci of evolution: how predictable is genetic evolution? *Evolution* 62: 2155-2177. doi: 10.1111/j.1558-5646.2008.00450.x
- Tirosch I, Barkai N 2007. Comparative analysis indicates regulatory neofunctionalization of yeast duplicates. *Genome Biol* 8: R50. doi: 10.1186/gb-2007-8-4-r50
- Tong AH, Evangelista M, Parsons AB, Xu H, Bader GD, Page N, Robinson M, Raghibizadeh S, Hogue CW, Bussey H, Andrews B, Tyers M, Boone C 2001. Systematic genetic analysis with ordered arrays of yeast deletion mutants. *Science* 294: 2364-2368. doi: 10.1126/science.1065810
- Trotter Eleanor W, Rand Jonathan D, Vickerstaff J, Grant Chris M 2008. The yeast Tsa1 peroxiredoxin is a ribosome-associated antioxidant. *Biochemical Journal* 412: 73-80. doi: 10.1042/bj20071634
- Wach A, Brachat A, Pohlmann R, Philippsen P 1994. New heterologous modules for classical or PCR-based gene disruptions in *Saccharomyces cerevisiae*. *Yeast* 10: 1793-1808.
- Wong C-M, Zhou Y, Ng RWM, Kung H-f, Jin D-Y 2002. Cooperation of Yeast Peroxiredoxins Tsa1p and Tsa2p in the Cellular Defense against Oxidative and Nitrosative Stress. *Journal of Biological Chemistry* 277: 5385-5394. doi: 10.1074/jbc.M106846200
- Yen K, Gitsham P, Wishart J, Oliver SG, Zhang N 2003. An improved tetO promoter replacement system for regulating the expression of yeast genes. *Yeast* 20: 1255-1262. doi: 10.1002/yea.1037
- Zhang J 2003. Evolution by gene duplication: an update. *Trends in Ecology & Evolution* 18: 292-298. doi: 10.1016/S0169-5347(03)00033-8

ANEXO H

Revisiones y trabajos publicados como autor secundario durante el doctorado

Genetic Redundancy. 2012. Ascencio D. and De Luna A. Encyclopedia of Systems Biology, Springer science Business and Media.

Molecular mechanisms of paralogous compensation and the robustness of cellular networks. Diss G., Ascencio D., DeLuna A., & Landry C. R. (2014). Journal of Experimental Zoology Part B: Molecular and Developmental Evolution, 322(7), 488-499.

Gene duplication and the evolution of moonlighting proteins.. Espinosa-Cantú A., Ascencio D., Barona-Gómez F., & DeLuna A. (2015). Frontiers in genetics, 6, 227.

Gene duplication can impart fragility, not robustness in the yeast protein interaction network. Diss G., Gagnon-Arsenault I., Dion-Coté A. M., Vignaud H., Ascencio D, Berger C.& Landr y C. R. (2017) Science, *accepted*