Genòmica evolutiva de la via de transducció de senyal de la insulina/TOR a insectes i vertebrats

David Álvarez Ponce

Genòmica evolutiva de la via de transducció de senyal de la insulina/TOR a insectes i vertebrats

Memòria presentada per DAVID ÁLVAREZ PONCE, inscrit al programa de Doctorat en Genètica, bienni 2005—2007, per optar al grau de Doctor per la Universitat de Barcelona. Aquest treball ha estat realitzat al Departament de Genètica de la Universitat de Barcelona sota la direcció dels Doctors JULIO ROZAS LIRAS i MONTSERRAT AGUADÉ PORRES.

Els directors L'autor

Julio Rozas Montserrat Aguadé David Álvarez

Barcelona, Juny de 2010

A mi madre y a mis amigos.

A la Irene, gran professora i amiga.

“One never notices what has been done,
one can only see what remains to be done”
— Marie Curie

Agraïments

Four Skinny Trees

They are the only ones who understand me. I am the only one who understands them. Four skinny trees with skinny necks and pointy elbows like mine. Four who do not belong here but are here. Four raggedy excuses planted by the city. From our room we can hear them, but Nenny just sleeps and doesn't appreciate these things.

Their strength is secret. They send ferocious roots beneath the ground. They grow up and they grown down and grab the earth between their hairy toes and bite the sky with violent teeth and never quit their anger. This is how they keep.

Let one forget his reason for being, they'd all droop like tulips in a glass, each with their arms around the other. Keep, keep, keep, trees say when I sleep. They teach.

When I am too sad and too skinny to keep keeping, when I am a tiny thing against so many bricks, then it is I look at the trees. When there is nothing left to look at on this street. Four who grew despite concrete. Four who reach and do not forget to reach. Four whose only reason is to be and be.

From *The House on Mango Street*
by Sandra Cisneros

Esta tesis trata, entre otras cosas, de cómo los genes no evolucionan de forma independiente, sino como piezas de las redes de las que forman parte. Lo mismo pasa con las personas: la evolución de cualquiera de nosotros depende enormemente de las personas que vamos encontrando a lo largo nuestro camino. Es por eso que **agradezco** a todas las personas que han estado a mi lado durante el proceso de realización de esta tesis, y que por lo tanto han contribuido, ya sea de forma directa o indirecta, a ella:

En primer lugar, a mis *supervisors*, **Julio** y **Montse**, por abrirme las puertas del mundo de la evolución molecular, por depositar su confianza en mi al darme la oportunidad de hacer esta tesis, por estar ahí, por inundar mi e-mail constantemente con artículos para leer, por intentar sacar lo mejor de mí, por tener paciencia con mis muchas excentricidades, y por muchas cosas más.

A **la Carme**, perquè les seves classes de claredat impecable eren tot un plaer i van fer-me decidir a fer el doctorat al grup, i també per guiar-me en les meves primeres passes pel laboratori. Al resto de miembros del **grupo de**

Genética Molecular Evolutiva que han coincidido conmigo, ya que de todos y cada uno de ellos he aprendido algo, con todos he compartido algún buen momento, y de todos me llevo algún buen recuerdo. En especial, a Eva e Inês, *las girls*, por ser mucho más que compañeras a lo largo de este viaje, y a mis compañeros de Bioinformática, por estar ahí cada vez que los he necesitado.

A **mi madre**, por quererme tanto, y por hacer todo lo que está en sus manos y más para que las cosas me vayan bien. Y a su pareja, **Enrique**, por estar ahí y por procurarme un techo bajo el que vivir durante estos últimos meses.

A todos **mis amigos**, por quererme y cuidarme tantísimo y por compartir tantos momentos (buenos y malos), viajes y risas conmigo, por ser las mejores ventanas por las que asomarse uno al mundo, porque todos sois irrepetibles e increíbles y hacéis el mundo mejor allí a donde vais, porque los días en que conocí a cada uno de vosotros fueron los días de más suerte de mi vida.

I de manera molt especial a **la Irene**, per un milió de coses, que per la seva modèstia de ben segur que no voldria que jo escrigués aquí. Però, per exemple, per contagiar-me el seu entusiasme per la Biologia i la recerca, per ser, en molts aspectes, un exemple a seguir, pels seus consells (sempre bons) i recolzament, i per regalar-me la seva amistat i confiança. Potser no estaria ara dipositant aquesta tesi si no t'hagués conegut.

A toda la gente del departamento con la que he tenido la oportunidad de intercambiar alguna historia, consejo, o simplemente un saludo o una sonrisa. I en especial al Professor **Lluís Serra**, per atreure la meva atenció primer envers la Genètica i després envers l'Evolució Molecular, en aquelles irrepetibles classes de segon de carrera.

To my English teachers too, **Griselda** and **Cristina**, 'cause, without your classes, I couldn't have written a word in English, and your work is also reflected in this thesis. You love your work, it's very easily seen. I already miss your classes!

A todas las personas que en algún momento se han parado frente a mí para enseñarme algo, cualquier cosa. A los escritores de los libros que me han mantenido más o menos cuerdo estos años.

Y, por supuesto, última pero sólo porque no sabes leer, a **mi gatita Txusma**, por quererme tan infinitamente, y por esperarme cada noche en casa con un millón de mimos.

Índex

1. INTRODUCCIÓ 1

1.1. Variabilitat genètica i detecció de la petjada de la selecció natural 3

1.1.1. Anàlisi filogenètica per màxima versemblança 5

1.2. Factors que afecten els nivells de limitació funcional a variar 8

1.2.1. Nivell i rang d'expressió gènica 10

1.2.2. Longitud de les proteïnes codificades 12

1.2.3. Biaix en l'ús de codons 12

1.3. Evolució de xarxes moleculars 14

1.3.1. Relació entre l'interactoma i l'evolució molecular dels seus components 16

1.3.2. Polaritat en el grau de limitació funcional al llarg de la via 18

1.3.3. Coeficients de control i nivell de limitació funcional 19

1.4. Via de transducció de senyal de la insulina/TOR 21

1.4.1. Via de la insulina/TOR a *Drosophila* 23

1.4.2. Via de la insulina/TOR als mamífers 26

1.5. Grups taxonòmics estudiats en aquesta tesi 29

1.5.1. Les espècies del gènere *Drosophila* 29

1.5.2. Els vertebrats 31

2. OBJECTIUS 35

3. ARTICLES 39

3.1. Article 1 - Network-level molecular evolutionary analysis of the insulin/TOR signal transduction pathway across 12 *Drosophila* genomes 41

3.1.1. Resum 41

3.1.2. Article 43

3.1.3. Material suplementari 53

3.1.4. Addenda 61

3.2. Article 2 - Comparative genomics of the vertebrate insulin/TOR signal transduction pathway genes: A network-level analysis of selective pressures along the pathway . 69
3.2.1. Resum 69
3.2.2. Article 71
3.2.3. Material suplementari 117

4. DISCUSSIÓ 141

4.1. Identificació dels gens implicats en la via de transducció de senyal de la insulina/TOR en els genomes de *Drosophila* i vertebrats 143
4.2. Impacte de la selecció natural sobre els gens de la via de la insulina/TOR 145
4.3. Relació entre l'estructura de la via de la insulina/TOR i els patrons d'evolució molecular dels seus components 146
4.3.1. Similitud en els patrons d'evolució molecular dels gens que codifiquen proteïnes que interactuen físicament 146
4.3.2. Polaritat en els nivells de limitació funcional al llarg de l'eix upstream/downstream de la via 148
4.3.3. Evolució dels gens que actuen als punts de ramificació de la via de la insulina/TOR 153
4.4. Comparació dels patrons d'evolució molecular de la via de la insulina a *Drosophila* i a vertebrats 154

5. CONCLUSIONS 157

6. BIBLIOGRAFIA 161

7. ANNEXOS 175

7.1. Annex 1 - Characterization of the insulin/TOR pathway genes in the *Pediculus humanus humanus* genome 177
7.2. Annex 2 - Informe dels directors de tesi 183

1. Introducció

1.1. Variabilitat genètica i detecció de la petjada de la selecció natural

La comparació de les seqüències de regions homòlogues del DNA (per exemple, un gen) entre diferents organismes normalment permet detectar diferències (alguns dels nucleòtids difereixen entre les seqüències comparades). Aquesta variabilitat entre seqüències homòlogues s'anomena polimorfisme o divergència, en funció de si els organismes comparats pertanyen o no a la mateixa espècie, respectivament. En els nivells i patrons d'aquesta variabilitat hi ha emmagatzemada una gran quantitat d'informació sobre la història evolutiva dels organismes comparats. Se'n pot extreure, per exemple, informació sobre les seves relacions evolutives (objecte d'estudi de la filogènia molecular), sobre la seva història demogràfica, o sobre les pressions selectives a les quals han estat sotmesos. Extreure tota aquesta informació de la variabilitat genètica és un dels objectius principals de la Genètica Evolutiva.

Atès que les seqüències homòlogues tenen, per definició, un origen comú (més o menys antic), totes les diferències que es poden detectar en comparar-les són el resultat de mutacions a nivell de DNA. Si bé les mutacions són esdeveniments que, en principi, tenen lloc de manera aleatòria, tot sovint els patrons de variabilitat observats en comparar seqüències homòlogues no es corresponen amb el que s'esperaria per atzar. Això és així perquè les diferències que s'observen en comparar seqüències homòlogues són un subconjunt de totes les mutacions que s'han produït al llarg de la seva història evolutiva: el destí de les variants resultants d'aquestes mutacions (per exemple, la seva fixació en les poblacions, o bé la seva pèrdua) està determinat per l'acció combinada de diferents forces evolutives, que comprenen la deriva genètica, la selecció natural i diferents esdeveniments demogràfics (per exemple, una variació en la grandària efectiva de la població).

En concret, la selecció natural actua eliminant aquelles mutacions que fan disminuir l'eficàcia biològica (definida com la capacitat de sobreviure i deixar descendència) dels individus que les porten (el que s'anomena selecció negativa o purificadora), o bé afavorint la fixació d'aquelles que la fan augmentar (selecció positiva o adaptativa). A més d'actuar sobre les diferències

esmentades (les que tenen un efecte sobre l'eficàcia biològica dels individus que les porten), la selecció natural també pot actuar sobre les variants neutres (és a dir, sense cap efecte sobre l'eficàcia biològica dels individus portadors de les diferents variants) a les quals, per la seva proximitat física en el genoma, hi estan lligades (el que s'anomena *linked selection*).

En la seva teoria neutralista de l'evolució molecular, Motoo Kimura (1968, 1983) va proposar que la majoria de la variabilitat genètica observada en comparar seqüències homòlogues seria selectivament neutra. D'acord amb aquesta teoria, la variabilitat genètica entre els organismes estaria majoritàriament governada per processos estocàstics (la mutació i la deriva genètica)[1].

Les prediccions de la teoria neutralista poden ser emprades com a model nul (el que s'anomena model neutre standard, SNM). Si es detecten diferències significatives entre els patrons de variabilitat observats i els esperats sota la hipòtesi nul·la, aquestes són generalment atribuïdes a l'efecte de la selecció natural. No obstant això, la petjada típica de la selecció natural sobre una regió del genoma pot ser confosa amb la que deixen determinats esdeveniments de tipus demogràfic (per exemple, la petjada típica de la selecció positiva recent coincideix amb la que deixa una expansió poblacional recent). Per tal de discriminar l'efecte de la selecció natural i dels factors demogràfics, es poden realitzar anàlisis multilocus: mentre que els esdeveniments demogràfics afecten a tots els *loci* del genoma, la selecció natural afecta a regions genòmiques concretes (per exemple, gens).

Els tests estadístics més emprats per detectar desviacions respecte del que s'espera segons la teoria neutralista es classifiquen, segons el tipus d'informació que fan servir, en: (1) els que fan servir només informació de polimorfisme; (2) els que es basen únicament en informació interespecífica (divergència); i (3) els que combinen informació dels dos tipus. Degut a la naturalesa de les dades emprades en aquesta tesi, només parlarem dels mètodes que utilitzen informació interespecífica, i en particular d'aquells que se centren en les regions codificadores dels gens. Aquestes regions estan formades per

[1]Aquesta teoria va ser posteriorment modificada per Tomoko Ohta i altres, donant lloc a la denominada teoria quasi neutralista de l'evolució molecular (Ohta 1973).

posicions sinònimes i no sinònimes. Només els canvis que afecten a les posicions no sinònimes tenen un efecte sobre la seqüència de les proteïnes codificades, amb la qual cosa els canvis a les posicions sinònimes es consideren, de manera general, selectivament neutres. Si totes les mutacions que tenen lloc a la regió codificadora d'un gen fossin selectivament neutres, esperaríem que, en comparar dues o més regions codificadores homòlogues, el nombre de substitucions no sinònimes per posició no sinònima (K_a o d_N) fos igual al nombre de substitucions sinònimes per posició sinònima (K_s o d_S), amb la qual cosa la relació $\omega = d_N/d_S$ seria de 1. Les desviacions de ω d'aquest valor s'interpreten com l'efecte de la selecció positiva (si $\omega > 1$) o de la purificadora ($\omega < 1$).

En absència de l'acció de la selecció positiva, ω pot ser considerada com una mesura del grau de limitació funcional a variar en la regió codificadora d'un gen: valors propers a zero són indicatius d'una forta selecció purificadora, mentre que valors propers a 1 serien el resultat d'una selecció purificadora més relaxada.

Pel que fa a la selecció positiva, invocar aquest tipus de selecció només quan s'observen valors de $\omega > 1$ és massa conservatiu: atès que la selecció positiva i negativa poden actuar de manera simultània sobre un mateix gen, si la selecció positiva actua només sobre uns pocs codons del gen, o bé en llinatges particulars, la seva petjada pot veure's emmascarada per la de la selecció purificadora. D'aquesta manera, és comú que, tot i que algunes posicions d'un gen hagin evolucionat de manera adaptativa, els seus valors globals de ω siguin menors que 1. Per aquest motiu, s'han desenvolupat mètodes d'anàlisi filogenètica per màxima versemblança, que permeten, entre d'altres coses, detectar la petjada de la selecció natural quan aquesta actua només sobre una porció dels codons d'un gen.

1.1.1. Anàlisi filogenètica per màxima versemblança

Els mètodes d'anàlisi filogenètica per màxima versemblança permeten, en primer lloc, obtenir estimes dels paràmetres d'un model evolutiu tot ajustant un conjunt de dades determinat (per exemple, un alineament múltiple de les seqüències codificadores d'un gen en diferents espècies) al model en qüestió. Durant els darrers anys s'ha desenvolupat tota una col·lecció de models,

cadascun dels quals té una sèrie de paràmetres i representa un escenari evolutiu determinat. L'ajust de les dades a un model consisteix en trobar la combinació de valors dels paràmetres que millor explica les dades observades (les anomenades estimes per màxima versemblança dels paràmetres del model). L'obtenció d'aquestes estimes representa un primer resultat d'aquest tipus d'anàlisi.

En segon lloc, a un mateix conjunt de dades es poden ajustar múltiples models, i de la comparació del grau d'ajust d'aquests models es poden fer inferències evolutives. En el context de la detecció de la selecció positiva, les dades es poden ajustar a dos models niats, el més complex dels quals (que representa un escenari amb selecció positiva) incorpora tots els paràmetres del més senzill més una sèrie de paràmetres addicionals que representen la classe de codons que evolucionen sota selecció positiva (típicament, la proporció que aquestes posicions representen respecte al total de codons, p_1, i el valor de ω d'aquesta classe, ω_1). La versemblança de tots dos models es compara mitjançant el test de raó de versemblança (Whelan i Goldman 1999), i si el model més complex s'ajusta significativament millor a les dades, es considera que ha actuat la selecció positiva. Una anàlisi semblant es pot aplicar per contrastar altres escenaris evolutius.

El programari més emprat per a aquest tipus d'anàlisi és l'anomenat *Phylogenetic Analysis by Maximum Likelihood* (PAML; Yang 1997). Entre els models implementats en aquest paquet de programes, destaquem els següents:

- M0: considera un únic valor de ω per a totes les branques de la filogènia i posicions del gen;
- *free-ratio* (FR): considera un valor de ω independent per a cadascuna de les branques de la filogènia;
- M1a: considera dues classes de codons: una que evoluciona amb $0 < \omega_0 < 1$, i una altra amb $\omega_1 = 1$;
- M2a: a més de les dues classes considerades al model M1a, permet una tercera classe de llocs que evolucionen sota selecció positiva ($\omega_s > 1$);
- M7: considera que els valors de ω es distribueixen d'acord amb una distribució beta amb $0 < \omega < 1$; i

- M8: permet, a més de la classe de llocs que evolucionen d'acord amb una distribució beta, una classe addicional de llocs que evolucionen sota selecció positiva ($\omega_s > 1$).

Es pot contrastar la presència d'una porció de llocs que evolucionen sota selecció positiva mitjançant la comparació dels models M1a i M2a (Wong et al. 2004), o bé dels models M7 i M8 (Yang et al. 2000). Un ajust significativament millor dels models M2a o M8 s'interpreta com a evidència de l'acció de la selecció positiva.

1.2. Factors que afecten els nivells de limitació funcional a variar

Després d'un esdeveniment d'especiació, els genomes de les espècies resultants comencen a acumular canvis de manera, en principi, independent. Cada regió genòmica divergeix a una taxa diferent, segons les diferents forces evolutives que hi actuen, cosa que converteix el genoma en un mosaic de conservació i divergència. Pel que fa a les regions codificadores dels gens, com s'ha mostrat en l'apartat anterior, en absència de selecció positiva el seu grau de limitació funcional a variar pot ésser avaluat d'acord amb la relació entre el grau de divergència sinònima i no sinònima ($\omega = d_N/d_S$).

Els valors de ω i d_N varien enormement (típicament al llarg de 2 o 3 ordres de magnitud) entre els diferents gens del genoma (veure, per exemple, Li et al. 1985). Un dels objectius de la genòmica evolutiva és determinar quins són els factors que tenen un efecte sobre els nivells de limitació funcional. La combinació de dades genòmiques amb conjunts de dades d'altres tipus ha permès identificar alguns factors que es correlacionen amb el grau de limitació funcional, i que per tant són factors potencialment implicats en la variabilitat dels nivells de limitació funcional entre els diferents gens. No obstant això, cal remarcar que el fet que un factor es correlacioni amb ω o d_N no implica una

Taula 1. Correlacions observades entre diferents factors genòmics amb el grau de limitació funcional i entre ells

	NP	C	GI	NE	CB	AP	E	PP
Nombre de paràlegs (NP)								
Connectivitat (C)[2]	+ +							
Interaccions genètiques (GI)	+ +	+						
Nivell d'expressió (NE)	+ + +	+ + +	-					
Biaix en l'ús de codons (CB)	ND	+ + +	ND	+ + +				
Abundància proteïna (AP)	ND	+ + +	ND	+ + +	+ + +			
Essencialitat (E)	+	+ + +	-	+ + +	+ + +	+ + +		
Propensió a la pèrdua (PP)	NS	- -	NS	- - -	ND	ND	- -	
Limitació funcional (LF)	- -	- - -	-	- - -	- - -	- - -	- -	+ + +

Els signes positius i negatius representen correlacions positives i negatives, respectivament. El nombre de signes representa el major o menor grau de correlació. ND, no determinada; NS, no significativa. Modificada de Koonin i Wolf 2006.

[2]Connectivitat: nombre de proteïnes amb les quals interactua una proteïna (veure apartat 1.3).

relació de causa-efecte entre ambdós factors: per exemple, el factor en qüestió podria estar correlacionat amb una altra variable, que al seu torn tingués un efecte directe sobre el grau de limitació funcional. De fet, molts dels factors que es correlacionen amb el grau de limitació funcional també es correlacionen entre ells (taula 1). Per tal de mitigar aquest problema, s'han fet servir tècniques d'anàlisi multivariant, com ara la correlació parcial o l'anàlisi de camins, que permeten l'anàlisi de les associacions entre les variables estudiades un cop descomptat l'efecte d'altres variables (figura 1).

Entre els factors que es correlacionen amb el grau de limitació funcional a variar es poden esmentar factors tan diversos com ara el nivell i el rang d'expressió gènica (Duret i Mouchiroud 2000; Pál et al. 2001; Subramanian i Kumar 2004), el biaix en l'ús de codons (Sharp 1991; Pál et al. 2001), la longitud de les proteïnes codificades (Subramanian i Kumar 2004), la categoria funcional

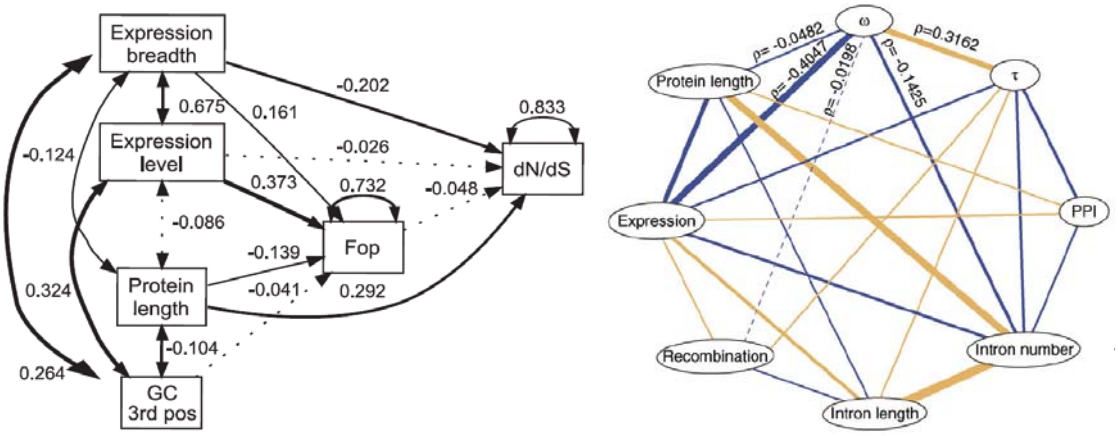

Figura 1. Anàlisi multivariant de l'efecte de diversos factors en la determinació del grau de limitació funcional. Esquerra: Anàlisi de camins per establir quins factors determinen el grau de limitació funcional a *Populus*. Les fletxes no puntejades i puntejades representen associacions significatives i no significatives, respectivament. Per a cada associació, s'especifica el coeficient del camí. Les fletxes que comencen i acaben en la mateixa variable representen la proporció de variabilitat no explicada pel model. *Expression breadth*, rang d'expressió gènica; *expression level*, nivell d'expressió; *protein length*, longitud de les proteïnes; *GC 3rd pos*, contingut de GC en les terceres posicions dels codons; *Fop*, freqüència de codons òptims. Extret de Ingvarsson 2007. Dreta: Anàlisi de correlacions parcials per establir quins factors determinen el grau de limitació funcional a *Drosophila*. Les línies blaves i taronges representen correlacions negatives i positives, respectivament. Les línies puntejades representen correlacions no significatives. Per a les correlacions que involucren ω, s'especifica el coeficient de correlació parcial (ρ). τ, biaix en l'expressió als diferents teixits; *PPI*, connectivitat; *Intron number*, nombre d'introns; *Intron length*, longitud dels introns; *Recombination*, taxa de recombinació; *Expression*, nivell d'expressió; *Protein length*, longitud de les proteïnes. Extret de Larracuente et al. 2008.

(veure, per exemple, Castillo-Davis et al. 2004), l'essencialitat[3] (Rocha i Danchin 2004) i la dispensabilitat (mesurada com la disminució en la taxa de creixement després de la pèrdua; Hirsh i Fraser 2001; Krylov et al. 2003; Yang et al. 2003; Drummond et al. 2006), el nombre i la longitud dels introns (Marais et al. 2005; Parmley et al. 2007), i la taxa de recombinació (Betancourt i Presgraves 2002; Marais et al. 2004; Zhang i Parsch 2005; Haddrill et al. 2007). A continuació es comenta l'efecte d'alguns d'aquests factors, en concret d'aquells que s'han emprat al llarg d'aquesta tesi, tot comentant els models que s'han proposat per explicar la seva relació amb el grau de limitació funcional. En moltes ocasions, però, els motius de les correlacions estan encara en discussió. A la figura 1 es mostren els resultats d'algunes anàlisis multivariants, que serveixen com a resum de les associacions esmentades. Per a una revisió més exhaustiva veure, per exemple, Pál et al. 2006 i Rocha 2006.

1.2.1. Nivell i rang d'expressió gènica

El nivell i el rang d'expressió gènica (nombre de teixits en què s'expressa un gen) són dos dels factors que d'una manera més clara es correlacionen amb el grau de limitació funcional (en organismes unicel·lulars, el nivell d'expressió explica aproximadament el 30–50% de la variabilitat en el nivell de limitació funcional; Pál et al. 2001; Drummond et al. 2005; Drummond et al. 2006), essent els gens que mostren un elevat nivell d'expressió, o que s'expressen en un elevat nombre de teixits, els que tendeixen a estar més fortament limitats a variar (Duret i Mouchiroud 2000; Pál et al. 2001; Subramanian i Kumar 2004).

El nivell i el rang de l'expressió gènica estan fortament correlacionats, i sembla que, en organismes pluricel·lulars, quan l'efecte de tots dos factors s'avalua simultàniament, només el rang d'expressió es correlaciona significativament amb el grau de limitació funcional (Pál et al. 2006; Ingvarsson 2007), la qual cosa indicaria que la correlació entre el nivell d'expressió i el grau de limitació funcional seria de fet un subproducte de l'associació entre el grau de limitació funcional i el rang d'expressió. No obstant això, la correlació entre el nivell d'expressió i el grau de limitació funcional és significativa quan s'avalua

[3]Un gen és essencial quan la seva deleció és letal.

de manera separada pels diferents teixits d'un organisme (veure, per exemple, Tuller et al. 2008), la qual cosa, juntament amb el fet que aquesta correlació s'observa en organismes unicel·lulars (on el concepte de rang d'expressió òbviament no es pot aplicar), valida el paper del nivell d'expressió com a factor implicat en la variabilitat dels nivells de limitació funcional.

La tendència dels gens que s'expressen en molts teixits a evolucionar sota una selecció purificadora més forta s'ha atribuït a que aquests gens (1) estarien implicats en un major nombre de processos bioquímics (Kuma et al. 1995); (2) han de funcionar en un ventall d'entorns bioquímics més divers, per exemple, interactuant amb un major nombre de molècules; o (3) les mutacions en aquests gens tindrien uns efectes pleiotròpics més importants (Duret i Mouchiroud 1999).

Pel que fa a la relació entre el nivell d'expressió i el grau de limitació funcional (independent del rang d'expressió), s'han proposat diferents models:

- Segons la *hipòtesi de la pèrdua de funció* (Rocha i Danchin 2004), cada molècula de proteïna faria una petita contribució a l'eficàcia biològica de l'organisme. D'aquesta manera, si es consideren dues mutacions que afecten la funció de dues proteïnes de manera equivalent, la mutació que afectés la proteïna més abundant tindria un efecte més gran sobre l'eficàcia biològica de l'organisme.
- Se sap que els gens amb un elevat nivell d'expressió estan fortament seleccionats per a l'ús dels codons sinònims que es tradueixen a una major velocitat o amb un major grau de precisió. Segons la *hipòtesi de l'eficiència traduccional* (Akashi 2001; Akashi 2003), les mutacions no sinònimes cap a codons subòptims serien lleugerament deletèries, especialment en aquests gens amb un elevat nivell d'expressió.
- Tots dos models presenten una sèrie d'inconsistències amb les observacions, i actualment el model més acceptat és la *hipòtesi de la robustesa traduccional* (Drummond et al. 2005). Se sap que la traducció és un procés amb una taxa d'error relativament elevada [uns 5 errors per cada 10.000 codons traduïts (Parker 1989), amb la qual cosa aproximadament el 19% de les proteïnes de llevats de longitud mitjana contindrien algun aminoàcid erroni]. Segons aquesta hipòtesi, les proteïnes diferirien en la seva tolerància a aquests errors: mentre que algunes serien capaces de plegar-se i funcionar

correctament tot i contenir aquests errors, d'altres tindrien més tendència a plegar-se erròniament, amb el conseqüent cost metabòlic i de toxicitat per la cèl·lula. La selecció purificadora tendiria a mantenir aquelles seqüències més robustes als errors en la traducció, especialment per a aquelles proteïnes que es tradueixen més freqüentment (Drummond et al. 2005).

1.2.2. Longitud de les proteïnes codificades

Els gens que codifiquen proteïnes més curtes evolucionen sota un major grau de limitació funcional (Subramanian i Kumar 2004; Lemos et al. 2005; Ingvarsson 2007). A *Populus*, aquest factor sembla ser el que té un efecte més important sobre el grau de limitació funcional (Ingvarsson 2007). Per explicar aquesta tendència, s'ha invocat la interferència en l'acció de la selecció natural a llocs propers, anomenada efecte Hill-Robertson (Comeron et al. 1999; McVean i Charlesworth 2000; Marais et al. 2005; Hill i Robertson 1966). En proteïnes llargues, la selecció natural actuaria de manera simultània sobre un elevat nombre de codons, tot reduint l'eficiència de la selecció natural.

1.2.3. Biaix en l'ús de codons

Els gens amb més limitació funcional presenten normalment un ús de codons més esbiaixat (Sharp 1991; Akashi 1994; Pál et al. 2001). S'ha suggerit que aquesta correlació resultaria del fet que els aminoàcids més importants per a la funció d'una proteïna estarien conservats al llarg de l'evolució, i també tendirien a estar codificats per codons òptims, que es tradueixen amb una major precisió (Akashi 1994). Una altra explicació seria que els gens en els quals hi hagués una relaxació de la selecció purificadora per tal de mantenir la seqüència d'aminoàcids també hi hauria una relaxació en la selecció per a l'ús de codons òptims (Comeron i Kreitman 1998). D'altra banda, el grau de biaix en l'ús de codons pot ésser considerat com una mesura de la taxa de traducció esperada d'una proteïna, amb la qual cosa, d'acord amb les hipòtesis de la *robustesa traduccional* i de la *pèrdua de funció*, s'espera que els gens que tenen un ús

fortament esbiaixat de codons estiguin fortament limitats a variar. També s'ha proposat una explicació de la correlació entre l'ús esbiaixat de codons i d_N (i ω) que invoca l'efecte de la selecció positiva: la fixació dels canvis no sinònims adaptatius interferiria amb la selecció feble per a l'ús de codons òptims a les posicions lligades (Betancourt i Presgraves 2002).

1.3. Evolució de xarxes moleculars

Les proteïnes no funcionen de manera aïllada: normalment es troben actuant com a peces de sistemes més complexes, com ara complexos proteics, vies metabòliques o de transducció de senyal, cèl·lules o organismes complets. D'aquesta manera, la funció d'una proteïna concreta només té sentit quan es considera la seva posició en una xarxa d'interaccions. En una xarxa metabòlica, un enzim normalment fa servir com a substrat un compost sintetitzat per un altre enzim, i produeix un producte que, al seu torn, serveix com a substrat d'un altre enzim. De manera similar, en una via de transducció de senyal l'estat funcional d'una proteïna (per exemple, estat actiu o inactiu) és modificat per altres components de la via (per exemple, mitjançant processos de fosforilació o desfosforilació) i, un cop activada, la proteïna pot, al seu torn, modificar la funció d'altres elements amb els quals interactua. Sota aquesta perspectiva, la funció d'un sistema bioquímic com ara una xarxa d'interaccions moleculars depèn, d'una manera complexa, de les propietats dels seus elements i de com aquests components estan connectats.

Si bé els projectes genoma han permès l'obtenció d'un catàleg més o menys complet de les proteïnes codificades pels genomes estudiats, se sap relativament poc de com tots aquests components interactuen entre si per generar la funció biològica. Caracteritzar les interaccions que tenen lloc entre les diferents entitats bioquímiques dels organismes, així com comprendre com la funció d'aquestes xarxes emergeix dels patrons d'aquestes interaccions i de les funcions d'aquests components són alguns dels objectius fonamentals de la Biologia de Sistemes.

L'estudi dels sistemes bioquímics des d'una perspectiva holística, a més de tenir un gran interès *per se*, obre la porta a un gran nombre d'aplicacions pràctiques. Per exemple, entendre el funcionament de les vies implicades en la síntesi d'un determinat compost d'interès pot ajudar a modificar-les amb l'objectiu d'incrementar-ne la producció. De la mateixa manera, entendre com funcionen les xarxes implicades en determinades patologies pot ajudar al disseny de fàrmacs.

Durant els darrers anys, s'ha anat acumulant informació sobre les xarxes d'interacció entre les proteïnes a diferents organismes (els anomenats *interactomes*), sobretot del llevat *Saccharomyces cerevisiae* (veure, per exemple, Yu et al. 2008), però també de l'ésser humà (veure, per exemple, Bossi i Lehner 2009) i de la mosca *Drosophila melanogaster* (Giot et al. 2003), entre d'altres organismes. Per exemple, Giot i col·laboradors van emprar el mètode de descoberta d'interaccions *yeast-two-hybrids* (Y2H) de manera massiva per tal de reconstruir l'interactoma de *D. melanogaster*, tot identificant un total de 20.405 interaccions entre 7.048 proteïnes (de 11.282 estudiades) (figura 2; Giot et al. 2003).

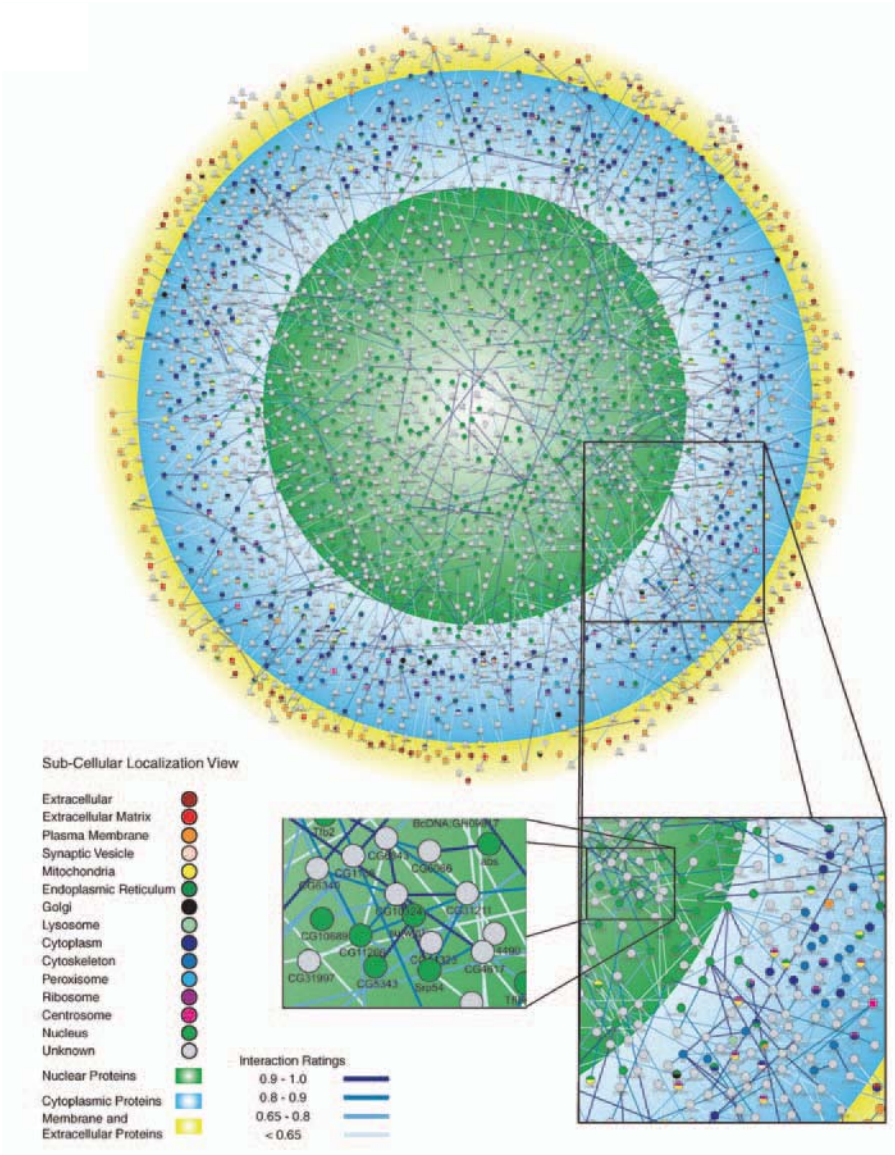

Figura 2. Interactoma de *D. melanogaster*. Extret de Giot et al. 2003.

Tot i que moltes de les interaccions proteïna–proteïna (PPIs) es coneixen en detall, la majoria de les PPIs conegudes s'han trobat mitjançant l'aplicació a gran escala de la tècnica Y2H. El baix nivell de coincidència entre els interactomes dels mateixos organismes reconstruïts per diferents laboratoris indica que, per una banda, encara quedarien moltes interaccions per descobrir, i que, per altra banda, moltes de les interaccions trobades en realitat representarien falsos positius (von Mering et al. 2002; Hart et al. 2006; Stumpf et al. 2008). Els interactomes dels què es disposa actualment s'han de considerar, doncs, com a esborranys.

Una manera de representar una xarxa d'interaccions moleculars és mitjançant un graf. Un graf és un objecte matemàtic format per una sèrie de nodes, que representen els elements (en aquest cas proteïnes), units per arestes o arcs, que representen relacions binàries (en aquest cas interaccions físiques).

1.3.1. Relació entre l'interactoma i l'evolució molecular dels seus components

La recent disponibilitat dels interactomes de diferents espècies ha permès estudiar la relació entre la seva estructura i els patrons d'evolució molecular dels seus components. Existeixen múltiples indicis que posen de manifest que la posició d'un gen en la xarxa d'interaccions té un efecte sobre la seva evolució, a diferents nivells (Cork i Purugganan 2004).

Un dels paràmetres que defineixen la posició d'una proteïna a l'interactoma és la *connectivitat* (el nombre de proteïnes amb les quals interactua). En tots els interactomes estudiats fins l'actualitat, la connectivitat segueix una distribució de llei potencial: la majoria de les proteïnes tenen una connectivitat baixa, mentre que unes poques (les anomenades *hubs*), estan involucrades en desenes o fins i tot centenars d'interaccions (Albert i Barabási 2002; Hahn i Kern 2005). Les proteïnes amb una elevada connectivitat ocupen una posició relativament central a la xarxa, mentre que les que tenen una connectivitat baixa se'n troben a la perifèria. Se sap que la connectivitat és un factor que es correlaciona amb el grau de limitació funcional de les proteïnes, essent els gens més connectats els que evolucionen sota un major grau de

limitació funcional (Fraser et al. 2002; Lemos et al. 2005). Aquesta associació s'ha atribuït al fet que, en les proteïnes amb un major grau de connectivitat, una major fracció dels aminoàcids estarien implicats en PPIs. De fet, Kim i col·laboradors van trobar que la fracció de la superfície de les proteïnes involucrada en PPIs es correlaciona amb ω millor que la connectivitat (Kim et al. 2006).

A més de la connectivitat, hi ha altres paràmetres derivats de la Teoria de Grafs que s'han emprat per caracteritzar la posició de les proteïnes a l'interactoma. Per exemple, dues mesures de centralitat àmpliament emprades són les anomenades *betweenness* i *closeness*. La *betweenness* és la freqüència amb la qual un node es troba en el camí més curt entre dos nodes de l'interactoma triats a l'atzar. Les proteïnes amb una elevada *betweenness* tindrien un paper molt important en controlar el flux d'informació al llarg de l'interactoma (Jeong et al. 2000; Wagner i Fell 2001), per exemple, connectant diferents mòduls de la xarxa. La *closeness* d'una proteïna, per la seva banda, és una mesura del nombre mitjà de passos necessari per arribar des d'aquesta fins tota la resta de proteïnes de la xarxa. Totes dues mesures de centralitat es correlacionen positivament amb el nivell de limitació funcional d'una manera més forta que la connectivitat (Hahn i Kern 2005).

Una altra evidència de que la posició dels elements en la xarxa d'interaccions té un efecte sobre l'evolució dels seus components és el fet que els gens que codifiquen proteïnes que interactuen entre sí presenten històries evolutives correlacionades. Per exemple, aquests gens evolucionen sota pressions selectives semblants (Fraser et al. 2002; Lemos et al. 2005). Això s'ha atribuït a la coevolució dels aminoàcids implicats en les PPIs (Fraser et al. 2002), o també a que les proteïnes que interactuen estarien sotmeses a nivells similars de selecció estabilitzadora (Lemos et al. 2005). De la mateixa manera, els gens que codifiquen proteïnes que interactuen també tendeixen a presentar nivells d'expressió semblants (Lemos et al. 2005), i també nivells de polimorfisme en l'expressió similars (Lemos et al. 2004). També s'ha proposat que els gens que codifiquen proteïnes que interactuen entre si tendirien a presentar històries evolutives semblants en quant a que es duplicarien i es perdrien en els mateixos llinatges (Fryxell 1996).

1.3.2. Polaritat en el grau de limitació funcional al llarg de la via

En una anàlisi de sis gens de la via de biosíntesi de les antocianines (figura 3), Rausher i col·laboradors van trobar una correlació entre el grau de divergència no sinònima i la posició que ocupen a la via (Rausher et al. 1999), essent els gens que actuen a la part inicial (*upstream*) de la via els que presenten un menor grau de divergència no sinònima. Anàlisis posteriors sobre un subconjunt d'aquests gens indiquen que aquesta distribució observada al llarg de la via seria deguda a diferències en la intensitat de la selecció purificadora que actua sobre aquests gens, i no a diferències en la taxa de mutació (Lu i Rausher 2003) ni en l'acció de la selecció positiva (Rausher et al. 2008).

Aquesta via presenta múltiples ramificacions, que donen lloc a la síntesi de diferents compostos derivats dels mateixos substrats inicials. Mentre que els gens situats a la part *upstream* de la via estarien implicats en la síntesi de tots aquests compostos, el nombre de compostos en la síntesi dels quals està implicat cadascun dels gens va disminuint a mesura que es progressa al llarg de la via. D'aquesta manera, el darrer gen de la via només està implicat en la síntesi

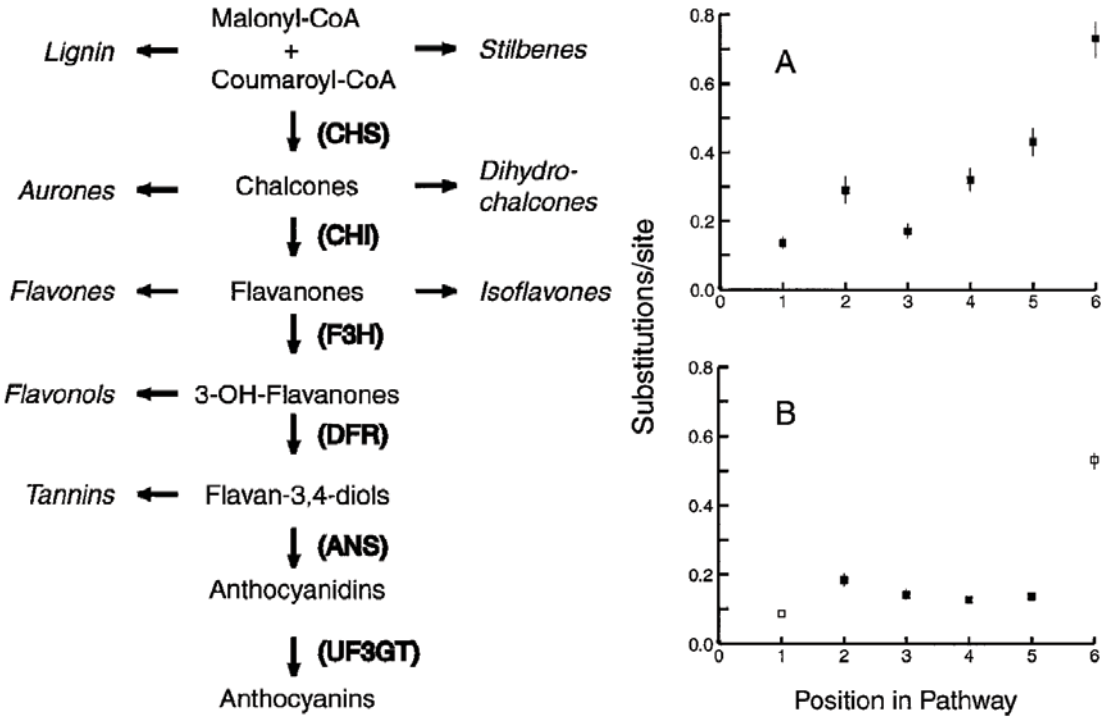

Figura 3. Anàlisi de la ruta de biosíntesi de les antocianines. Esquerra: Estructura de la via. Entre parèntesi es representen els enzims implicats. Dreta: Correlació entre la posició a la via dels gens de la ruta de biosíntesi de les antocianines i el grau de divergència no sinònima, calculat com a la mitjana de la comparació *Zea mays–Ipomoea purpurea* i *Z. mays–Antirrhinum majus* (A), o bé a partir de la comparació *I. purpurea–A. majus*. Extret de Rausher et al. 1999.

d'antocianines. Aquesta polaritat al llarg de la via en el nombre de compostos en la síntesi dels quals està implicat cada gen s'ha emprat per explicar la polaritat en el grau de limitació funcional observada: les mutacions que tenen lloc a la part *upstream* de la via afectarien un major nombre de fenotips, i per tant tindrien uns majors efectes pleiotròpics que les mutacions que afecten als gens de la part final (*downstream*).

S'ha trobat una distribució similar dels nivells de limitació funcional al llarg d'altres vies biosintètiques en plantes, com ara la ruta de l'isoprè, la dels terpenoides i la dels carotenoides (Sharkey et al. 2005; Livingstone i Anderson 2009; Ramsay et al. 2009), i també a la via de senyalització Ras de *Drosophila* (Riley et al. 2003). No obstant això, en altres vies no s'observa una correlació entre la posició dels gens i el grau de limitació funcional de les proteïnes codificades, la qual cosa mostra que aquest patró no seria universal (Olsen et al. 2002; Jovelin et al. 2009; Yang et al. 2009).

1.3.3. Coeficients de control i nivell de limitació funcional

Els enzims d'una via poden contribuir de manera diferencial a la seva funció (i, per tant, als fenotips en els quals la via està implicada). La funció de la via pot ser molt sensible a les característiques cinètiques de determinats enzims, mentre que pot ser més o menys independent de la funció d'altres enzims. La dependència de la funció global de la via de les característiques de cadascun dels seus components es pot mesurar mitjançant els coeficients de control (Kacser i Burns 1973).

S'espera que els gens que tenen coeficients de control elevats (és a dir, amb una elevada influència sobre la funció de la via), estiguin sotmesos a pressions selectives més fortes que els gens amb coeficients de control més baixos, que evolucionarien sota pressions selectives més febles (Hartl et al. 1985; Eanes 1999; Watt i Dean 2000). Per exemple, s'espera que els enzims que actuen en punts de ramificació de la via tinguin un efecte relativament important sobre el flux (ja sigui d'informació o de metabòlits) al llarg de la via, i que per tant tinguin uns elevats coeficients de control (LaPorte et al. 1984; Stephanopoulos i Vallino 1991). D'acord amb aquesta visió, Flowers i

col·laboradors van trobar la petjada de la selecció positiva en 5 gens que actuen en punts de ramificació de les vies del catabolisme de la glucosa a *Drosophila* (figura 4; Flowers et al. 2007).

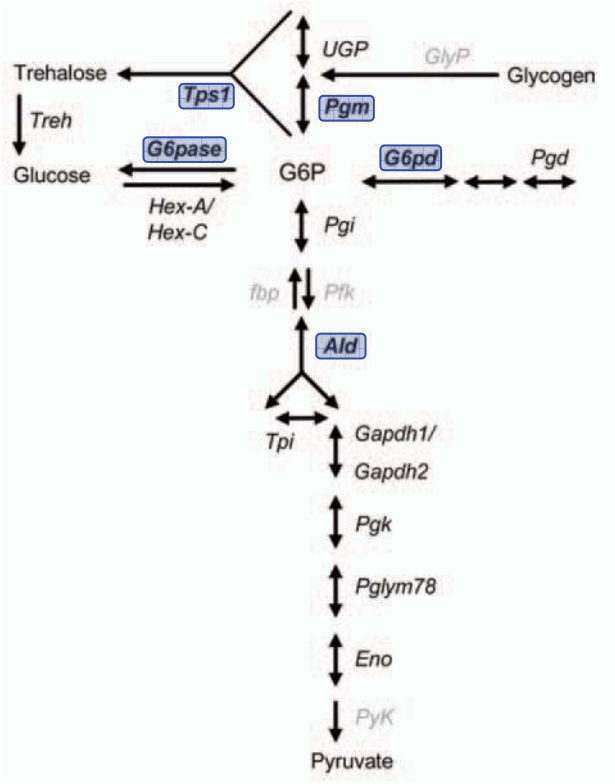

Figura 4. Punt de ramificació de la glucosa-6-fosfat (G6P) i vies adjacents a *Drosophila*. En blau es remarquen els gens que han evolucionat de manera adaptativa. Modificat de Flowers et al. 2007.

1.4. Via de transducció de senyal de la insulina/TOR

La via de la insulina/TOR té un paper clau en el control de l'homeòstasi dels animals. Està implicada en processos tan fonamentals i diversos com el metabolisme energètic i el creixement, tant a nivell cel·lular com d'organisme, la reproducció i l'envelliment.

La insulina és una hormona proteica que se secreta en resposta a, entre altres factors, alts nivells de glúcids al sistema circulatori (per exemple, després de la ingesta d'aliments). La interacció de la insulina amb el receptor de la insulina (ubicat a la membrana cel·lular dels teixits diana) promou la incorporació de glúcids a la cèl·lula, la síntesi de glicogen, proteïnes i lípids, i el creixement i la proliferació cel·lular.

La desregulació de determinants elements d'aquesta via està implicada en patologies com ara la resistència a la insulina, la diabetis, l'obesitat i el càncer, la qual cosa fa que aquesta via sigui de gran interès en biomedicina. Descoberta al 1921, la insulina va ser el primer pèptid que va ser administrat terapèuticament[4], i també el primer a ésser seqüenciat completament (Sanger i Tuppy 1951a,b; Sanger i Thompson 1953a,b)[5]. Si es fa una cerca a PubMed amb el terme "insulin", es recuperen un total de 260.570 articles (a 28/05/2010), dels quals 13.938 van ser publicats durant el 2009. Tot i els grans esforços dedicats a l'estudi d'aquesta via, la comprensió actual del seu funcionament encara dista molt d'ésser total, i amb freqüència se'n van descrivint nous components i interaccions, tant entre els components de la via com entre aquests i altres vies.

La funció i l'arquitectura de la xarxa d'interaccions que permet el flux del senyal des del receptor de la insulina fins els efectors finals de la via ha estat

[4]La descoberta de la insulina, inicialment atribuïda a Fredrick Banting, Charles Best, John James Richard Macleod i James Collip (Banting et al. 1922), va valdre el Premi Nobel de Medecina (1923) a Banting i a Macleod, que van compartir-ne la part econòmica amb els altres dos investigadors. El grup va vendre la patent d'aquesta descoberta a la Universitat de Toronto per 1 dòlar canadenc. Posteriorment, però, aquesta descoberta s'ha atribuït a l'investigador romanès Nicolae Paulescu, que havia realitzat un treball semblant durant el 1916, que va publicar al 1921 (vuit mesos abans que els investigadors canadencs) en revistes europees. Sembla ser que els investigadors canadencs no van entendre correctament aquests treballs, publicats en francès, i els van citar erròniament.

[5]Frederick Sanger va rebre el seu primer Premi Nobel de Química (1955) pels seus treballs sobre l'estructura de les proteïnes, i en especial sobre la de la insulina.

caracteritzada en profunditat en diferents organismes [en especial a *D. melanogaster* (figura S1 de l'article 1), *Caenorhabditis elegans* i als mamífers]. La via està essencialment conservada tant a nivell fisiològic com molecular al llarg de tot el regne animal (inclús les esponges semblen tenir un receptor de la insulina; Skorokhod et al. 1999), la qual cosa reflexa la importància de les funcions que regula. Un exemple del grau de conservació de la via entre artròpodes i vertebrats és el fet que la insulina humana pot estimular el receptor de la insulina de *D. melanogaster* (Fernandez et al. 1995; Yamaguchi et al. 1995), i un extracte de proteïnes de *D. melanogaster* té bioactivitat insulina sobre el ratolí (Meneses i De Los Angeles Ortiz 1975). A més, les mutacions que afecten els gens implicats a la via de la insulina tenen efectes similars a mamífers i a *Drosophila* [entre d'altres, reducció de la mida corporal (figura 5), elevats nivells de glúcids en circulació i increment de la durada de la vida].

Figura 5. Efecte de la via de la insulina/TOR sobre la mida corporal. Esquerra: Comparació d'un individu *chico*$^{-/+}$ (a dalt) amb un de *chico*$^{-/-}$ (a baix) de *D. melanogaster*; extret de Oldham et al. 2000. Dreta: Comparació d'un ratolí *IRS1*$^{-/-}$ (esquerra) amb un de *IRS1*$^{-/+}$ (dreta); extret de LeRoith et al. 2000. En tots dos casos la reducció a la mida corporal és del 40–50%. Nota: *IRS1* és un dels homòlegs de *chico* al genoma del ratolí.

En aquest apartat es revisen els mecanismes moleculars que permeten la transducció del senyal des del receptor de la insulina fins la resta d'elements de la via a *Drosophila*, i després s'assenyalen algunes diferències respecte als mamífers. Per a una descripció més detallada de la via de la insulina, veure, per exemple Oldham i Hafen 2003, LeRoith et al. 2004, Taguchi i White 2008 i Teleman 2010. Veure la figura de la via de la insulina a *Drosophila* a l'article 1 (figura S1).

1.4.1. Via de la insulina/TOR a *Drosophila*

Al genoma de *D. melanogaster* s'han identificat un total de 7 gens que codifiquen les anomenades *insulin-like peptides* (ILPs), homòlogues a la insulina i els *insulin-like growth factors* (IGFs) dels vertebrats. Aquests gens s'expressen a diferents òrgans i teixits: 7 cèl·lules neurosecretòries a cada hemisferi del cervell (*dilp1*, *2*, *3* i *5*; Brogiolo et al. 2001; Ikeya et al. 2002; Broughton et al. 2005), l'intestí mitjà (*dilp4*, *5* i *6*), el cordó nerviós ventral (*dilp7*; Brogiolo et al. 2001), les cèl·lules fol·liculars dels ovaris (*dilp5*; Brogiolo et al. 2001), el mesoderm embrionari (*dilp4*; Brogiolo et al. 2001), els discs imaginals i les glàndules salivars (*dilp2*; Brogiolo et al. 2001; Ikeya et al. 2002).

Un cop alliberada al sistema circulatori, la insulina arriba als teixits diana, on interactua amb el receptor de la insulina (InR, receptor situat a la membrana cel·lular). La interacció de l'hormona amb el domini extracel·lular del receptor promou l'activitat cinasa del domini intracel·lular, que d'aquesta manera pateix autofosforilació. En el seu estat fosforilat, el receptor de la insulina és capaç de reclutar la proteïna adaptadora Chico[6] (Yenush et al. 1996). Chico, al seu torn, és capaç de reclutar el complex PI3K [format per una subunitat reguladora, p60 (Weinkove et al. 1997), que és la que interactua directament amb Chico, més una subunitat catalítica, p110 (Leevers et al. 1996)]. A diferència del receptor de la insulina dels vertebrats, el de *Drosophila* té una extensió C-terminal d'uns 368 aminoàcids que conté tres llocs d'interacció amb PI3K, i per tant és capaç de reclutar aquesta proteïna de manera independent de Chico (Fernandez et al.

[6]Veure llistat dels gens que codifiquen les proteïnes esmentades en aquest apartat a la taula S1 de l'article 1.

1995; Ruan et al. 1995). Aquesta activació directa de PI3K, però, no és suficient per suplir completament la senyalització a través de Chico (Oldham i Hafen 2003). A més, aquesta extensió C-terminal és eliminada proteolíticament en alguns teixits (Fernandez et al. 1995).

Un cop a la membrana, el complex PI3K és capaç de catalitzar la síntesi del lípid de membrana fosfatidilinositol (3,4,5)-trisfosfat (PIP_3) tot fosforilant el fosfatidilinositol (4,5)-bisfosfat (PIP_2) a la posició D3 de l'anell inositol. El PIP_3 és un missatger secundari capaç de reclutar cap a la cara interna de la membrana cel·lular les proteïnes que contenen un domini *pleckstrin homology* (PH), com ara Melted, PDK1, PKB i la pròpia Chico. El cúmul de PIP_3 a la membrana (i, per tant, la localització a la membrana d'aquestes proteïnes) és transitori, ja que aquest és convertit en PIP_2 per la fostatasa de fosfoinosítids PTEN (per a una revisió, veure Vinciguerra i Foti 2006).

Quan, degut a l'acció de la insulina, les proteïnes PDK1 i PKB es troben a la membrana, aquestes poden interactuar físicament, la qual cosa permet que PDK1 fosforili PKB (per a una revisió, veure Franke 2008). En el seu estat fosforilat, PKB és capaç de fosforilar (i, per tant, inhibir) tres proteïnes inhibitòries de la senyalització de la insulina: Shaggy (Papadopoulou et al. 2004), dFOXO (Kramer et al. 2003; Puig et al. 2003) i Tsc2 (Potter et al. 2002).

Un cop activada, Shaggy inhibeix les proteïnes dMyc (Galletti et al. 2009), eIF2B-ε i la glicogen sintasa. Per tant, la inhibició de Shaggy per part de PKB activa aquestes tres proteïnes. El factor de transcripció dMyc promou la transcripció d'un gran nombre de gens (s'ha proposat que al voltant del 15% dels gens de *D. melanogaster*; Orian et al. 2003), entre els quals hi ha gens implicats en la gènesi dels ribosomes (Grewal et al. 2005; Teleman et al. 2008). La fosforilació de dMyc per part de Shaggy en promou la degradació (Galletti et al. 2009). La proteïna eIF2B-ε és la subunitat catalítica del complex eIF2B (*eukaryotic initiation factor 2B*), la funció del qual és necessària per tal de reclutar el Met-$tRNA_i$ cap al ribosoma.

dFOXO és un factor de transcripció que, en el seu estat activat, és translocat al nucli on promou la transcripció de, entre d'altres, els gens que codifiquen el receptor de la insulina i la proteïna d4E-BP (veure més avall). En resposta a la insulina, dFOXO pot ser inactivat mitjançant dos mecanismes. En

primer lloc, pot ser fosforilat per PKB, i un cop fosforilat és segrestat al citoplasma per proteïnes de la família 14-3-3, la qual cosa n'impedeix la translocació al nucli (Puig et al. 2003; Van Der Heide et al. 2004). En segon lloc, la proteïna Melted, que, en resposta a la insulina, es localitza a la membrana, s'uneix a dFOXO tot impedint-ne la localització nuclear (Teleman et al. 2005), i al mateix temps facilitant-ne la interacció amb PKB. L'activació del receptor de la insulina per part de dFOXO (que, al seu torn, és inactivat per la via de la insulina) s'interpreta com un bucle de retroacció negativa que limita el nivell d'activació de la via de la insulina (Puig et al. 2003).

El complex TSC consta de dues subunitats, Tsc1 i Tsc2, que s'uneixen mitjançant els seus dominis *coiled-coil*. Quan aquest complex està assemblat, és funcional, i promou l'activitat GTPasa intrínseca de Rheb (que impedeix la formació del complex Rheb-GTP, la forma activa de Rheb). La via de la insulina pot inhibir el complex TSC (i per tant permetre l'existència del complex Rheb-GTP) de dues maneres. En primer lloc, PKB fosforila Tsc2 (Potter et al. 2002), tot promovent la dissociació del complex. En segon lloc, Melted s'uneix a la membrana cel·lular i a Tsc1 (Teleman et al. 2005; Cai et al. 2006), la qual cosa facilita la interacció del complex amb PKB.

La forma activa de Rheb (Rheb—GTP) s'uneix a la proteïna TOR, tot activant-la. TOR també pot ésser activada directament per aminoàcids, de manera independent de la via de la insulina (revisat a Teleman 2010). TOR és la subunitat catalítica de dos complexes diferents: TORC1 (implicat en la regulació del creixement cel·lular i en la síntesi de proteïnes) i TORC2 (menys caracteritzat). La funció de TORC1 té lloc, principalment, mitjançant la fosforilació de les proteïnes d4E-BP (que implica la seva inhibició) i S6K (que n'implica l'activació) (Miron et al. 2003).

En la seva forma activa, la proteïna d4E-BP (*Drosophila 4E binding protein*) s'uneix a les proteïnes eIF-4E (a *D. melanogaster* se n'han descrit 8 isoformes, de les quals no totes serien funcionals, codificades per 7 gens paràlegs; Hernandez et al. 2005), tot impedint-ne la funció. L'activació de TORC1 per part de la via de la insulina promou, doncs, l'alliberament de les proteïnes eIF-4E, i per tant la seva funció. Les proteïnes eIF4E (*elongation*

initiation factors 4E) estan implicades en el reclutament del ribosoma cap a la part 5' dels mRNAs (mitjançant la interacció amb l'estructura CAP, m^7Gppp).

A més de ser fosforilada per TORC1, la proteïna S6K necessita ésser fosforilada en una altra posició per part de PDK1 per tal d'activar-se completament (Chou i Blenis 1995; Dufner i Thomas 1999; Avruch et al. 2001). Un cop activada, S6K (S6 *kinase*) fosforila (i activa) la proteïna ribosomal RpS6 (Miron et al. 2003).

Al contrari que TORC1, el complex TORC2 és inhibit per Rheb (i, per tant, per la via de la insulina) (Yang et al. 2006). Aquest complex és capaç de fosforilar PKB, amb un efecte activador, la qual cosa s'interpreta com un segon bucle de retroacció negativa: en condicions de baixos nivells d'activació de la via de la insulina, s'activa el complex TORC2, que al seu torn activa PKB (Yang et al. 2006).

1.4.2. Via de la insulina/TOR als mamífers

En la seva forma madura, la insulina humana és una proteïna formada per dos pèptids, A i B, de 21 i 30 aminoàcids respectivament. Aquesta proteïna se sintetitza a les cèl·lules ß dels illots de Langerhans (pàncrees), on s'emmagatzema en vesícules de secreció i se secreta en resposta a alts nivells de glucosa en sang (entre d'altres factors).

Una diferència fonamental entre la via de la insulina a *Drosophila* i a mamífers és el nombre de gens implicats. A *D. melanogaster*, la majoria dels gens que codifiquen les proteïnes esmentades a l'apartat anterior són de còpia única, la qual cosa n'ha facilitat l'estudi. Als vertebrats, per contra, moltes d'aquestes proteïnes són codificades per múltiples gens paràlegs[7], la qual cosa augmenta la complexitat de la via. Tot i que la funció acostuma a estar conservada entre els gens de mamífers i de *Drosophila*, sovint els estudis funcionals s'han centrat en un subconjunt de les còpies presents als genomes de mamífers, i per tant el coneixement actual sobre la funció dels diferents gens paràlegs és parcial.

[7]Per a un llistat d'aquests gens, veure la taula S1 de l'article 2.

Per exemple, el propi gen de la insulina compta amb dos paràlegs propers en el genoma humà, que codifiquen els *insulin-like growth factors* (IGF) 1 i 2. Aquestes tres hormones tenen efectes solapats però diferents: mentres que la insulina té efectes sobre el metabolisme, els IGFs (que se secreten majoritàriament al fetge en resposta a l'hormona del creixement, GH) controlen principalment el creixement. El gen que codifica el receptor de la insulina també té dos paràlegs propers al genoma humà, que codifiquen el receptor de IGF1 (IGF1R), i el *insulin receptor-related receptor* (IRR). El receptor de la insulina és activat per IGF2 durant el període prenatal i per la insulina durant el període postnatal, mentre que IGF1R és activat per IGF1 i 2 al període prenatal i per IGF1 al període postnatal (revisat per Nakae et al. 2001). IRR, per contra, no seria activat per cap dels tres ligands (Jui et al. 1994). Tant el receptor de la insulina con el de IGF1 són capaços d'unir-se a les proteïnes IRS (homòlogues de Chico), tot desencadenant l'activació de la via de la insulina/TOR.

Tot i que l'estructura de la via de la insulina està essencialment conservada entre *Drosophila* i els mamífers (de fet, totes les interaccions descrites a l'apartat anterior s'han confirmat també als mamífers[8]), s'han trobat algunes diferències[9] (per a una revisió, veure Teleman 2010). Algunes d'aquestes diferències representen interaccions descobertes en un sistema que encara no s'han estudiat a l'altre, mentre que d'altres han pogut ser confirmades. Per exemple, als mamífers la incorporació de glucosa als teixits sensibles a la insulina està controlada pel transportador GLUT4, que és translocat a la membrana en resposta a la insulina. Tot i que a *D. melanogaster* existeix un ortòleg de GLUT4, sembla que la insulina no afecta la incorporació de glucosa a les cèl·lules en aquest organisme (Ceddia et al. 2003). Una altra diferència és la manera en què aquesta via està connectada amb altres vies; per exemple, a mamífers la via de la insulina està fortament inhibida per la via del receptor de glucocorticoides (van Raalte et al. 2009), mentre que a *Drosophila* no hi ha un ortòleg obvi d'aquest receptor (King-Jones i Thummel 2005). També sembla haver-hi diferències respecte a la connexió de la via de la insulina amb la via Ras: mentre que als mamífers l'activació de la via Ras és un important

[8]Amb excepció de la activació de dels gens que codifiquen les 4E-BPs per part dels factors de transcripció FOXO. Si bé aquesta interacció va ser descrita en un article (Southgate et al. 2007), aquest va ser retractat.
[9]Comparar la figura 2 de l'article 1 i la figura 2 de l'article 2.

mediador dels efectes mitogènics de la via de la insulina (Margolis i Skolnik 1994; Ogawa et al. 1998), a *Drosophila* aquest efecte de Ras sobre els efectes de la insulina sembla ser només modulatori (Orme et al. 2006).

1.5. Grups taxonòmics estudiats en aquesta tesi

En aquesta tesi s'ha estudiat l'evolució molecular de via de la insulina/TOR de manera separada en dos grups d'organismes diferents: 12 espècies del gènere *Drosophila*, i 6 de vertebrats.

1.5.1. Les espècies del gènere *Drosophila*

De les aproximadament 3000 espècies descrites de dípters de la família *Drosophilidae*, més de 2000 pertanyen al gènere *Drosophila* (Markow i O'Grady 2005), la majoria de les quals s'agrupen en els subgèneres *Sophophora* (~1100 espècies), i *Drosophila* (~330 espècies). Si bé la majoria d'espècies té una longitud de 2–4 mm, algunes són més grans que la mosca domèstica. Diverses espècies d'aquest gènere, i sobretot *D. melanogaster*, han esdevingut organismes model en àrees de la Biologia tan diverses com la Biologia Molecular, la Biologia del Desenvolupament, l'Ecologia, la Biologia de Poblacions, la Genètica i l'Evolució. Això es deu, en part, a motius històrics, però sobretot a determinades característiques pràctiques que fan que aquests organismes siguin idonis per treballar-hi al laboratori, com ara la facilitat amb què es poden capturar i mantenir, el curt temps de generació i l'elevat nombre de descendents per individu.

L'any 2000 es va publicar la fracció eucromàtica del genoma de de *D. melanogaster* (Adams et al. 2000), el qual està extraordinàriament ben ensamblat i anotat. Amb la publicació, al 2005, del genoma de *D. pseudoobscura* (Richards et al. 2005) es va disposar, per primera vegada a la història, de les seqüències genòmiques de dues espècies eucariotes pluricel·lulars del mateix gènere. L'any 2007 es van publicar les seqüències genòmiques de 10 espècies més del mateix gènere (*Drosophila* 12 Genomes Consortium 2007), amb diferents temps de divergència (figura 6): *D. simulans*, *D. sechellia*, *D. yakuba*, *D. erecta*, *D. ananassae*, *D. persimilis*, *D. willistoni*, *D. mojavensis*, *D. virilis* i *D. grimshawi*. La disposició d'un total de 12 genomes del mateix gènere, combinada amb l'excepcionalment bona qualitat de l'ensamblat

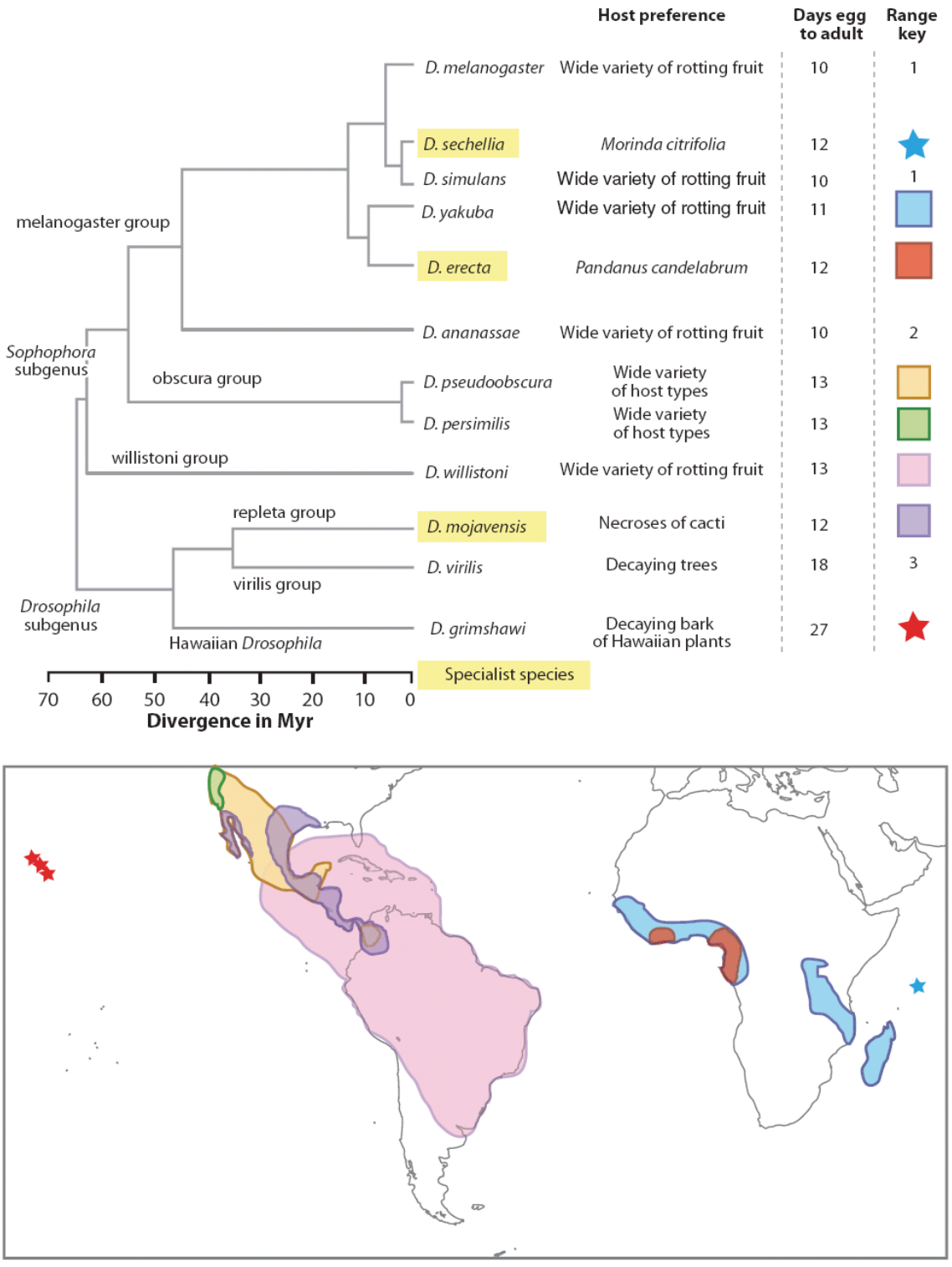

Figura 6. Dotze espècies de *Drosophila* amb genoma seqüenciat. A dalt: filogènia de les 12 espècies. Per a cada espècie, es mostra l'hoste preferit per a l'ovoposició, el temps de desenvolupament i una clau del rang de distribució geogràfica. A baix: Distribució geogràfica de les espècies, excepte les que estan marcades amb un nombre (1, espècies cosmopolites; 2, espècie cosmopolita, altament freqüent a Àsia i al Pacífic; 3, espècie holàrtica). Modificada de Singh et al. 2009.

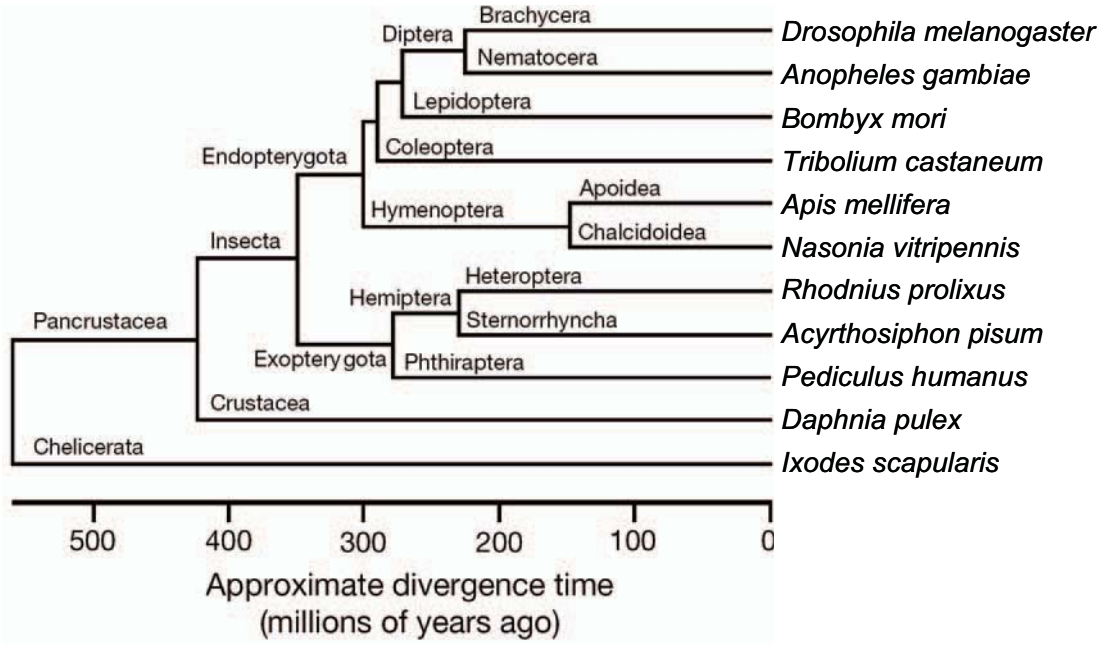

Figura 7. Filogènia dels artròpodes. Modificada de The Honeybee Genome Sequencing Consortium 2006.

i l'anotació del genoma de *D. melanogaster*, va obrir la possibilitat a anàlisis evolutives fins aleshores impensables. A més, en els darrers anys també s'han seqüenciat, o estan en vies de seqüenciació, els genomes d'una sèrie d'espècies d'artròpodes (figura 7).

Aquestes 12 espècies, que van divergir fa uns 40–60 milions d'anys (Russo et al. 1995; Tamura et al. 2004), varien enormement en el seu rang de distribució geogràfica (des d'espècies cosmopolites, com ara *D. melanogaster* i *D. simulans*, fins a endemismes d'illes, com *D. sechellia* i *D. grimshawi*), alimentació (amb espècies generalistes i d'altres que s'alimenten únicament dels fruits d'una espècie concreta), comportament, mida, grandària del genoma, etc. (veure, per exemple, Markow i O'Grady 2007) (figura 6).

1.5.2. Els vertebrats

Els vertebrats són un subfílum de cordats proveïts d'un teixit ossi format per vèrtebres que protegeixen el sistema nerviós central, i d'una caixa cranial que protegeix l'encèfal. D'acord amb la classificació tradicional, s'agrupen en

set classes: tres classes de peixos (unes 31.500 espècies) i les classes *Amphibia* (unes 6350 espècies), *Reptilia* (unes 8200 espècies), *Aves* (unes 10.000 espècies) i *Mammalia* (unes 5300 espècies). Els mamífers al seu torn, es classifiquen en tres subclasses: els monotremes (*Prototheria*, ponedors d'ous, 5 espècies), els marsupials (*Metatheria*, ~330 espècies) i els placentaris (*Eutheria*, ~5000 espècies). Es creu que els vertebrats es van originar fa uns 500–600 milions d'anys (Genome 10K Community of Scientists 2009), durant l'explosió càmbrica, i que els mamífers van aparèixer fa uns 225 milions d'anys.

El genoma humà, el primer esborrany del qual es va publicar l'any 2001 (International Human Genome Sequencing Consortium 2001, 2004; Venter et al. 2001), va ser el primer genoma de vertebrat a ésser seqüenciat. Des d'aleshores, els avenços en les tècniques de seqüenciació han permès l'obtenció de les seqüències genòmiques de múltiples espècies de vertebrats: a la base de dades Ensembl versió 57 (Hubbard et al. 2009) són accessibles els genomes de 6

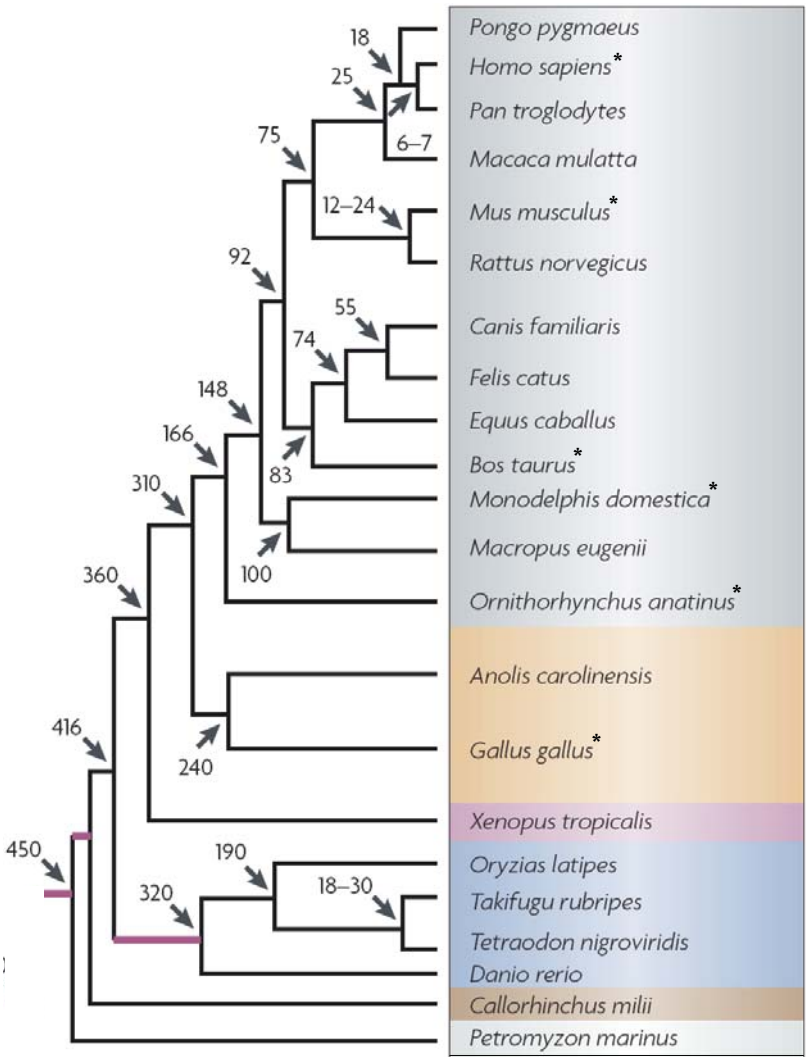

Figura 8. Filogènia dels vertebrats. Les espècies estudiades en aquesta tesi es marquen amb un asterisc. Els nombres representen els temps de divergència estimats (en milions d'anys). En lila es representen les branques internes a les quals s'han proposat processos de duplicació del genoma complet. Modificat de Ponting 2008.

peixos, un amfibi, un rèptil, tres ocells i 36 mamífers, en diferents estadis i amb diferents graus de cobertura. Recentment, s'ha proposat un projecte que té com a objectiu obtenir la seqüència genòmica de 10.000 espècies de vertebrats (aproximadament una espècie per gènere; Genome 10K Community of Scientists 2009).

En aquesta tesi s'ha treballat amb els genomes del pollastre (*Gallus gallus*; International Chicken Genome Sequencing Consortium 2004) i de cinc mamífers: 3 placentaris [l'ésser humà (*Homo sapiens*; International Human Genome Sequencing Consortium 2001, 2004; Venter et al. 2001), el ratolí (*Mus musculus*; Mouse Genome Sequencing Consortium 2002) i la vaca (*Bos taurus*; The Bovine Genome Sequencing and Analysis Consortium 2009)], un marsupial (l'opòssum cuacurt gris, *Monodelphis domestica*; Mikkelsen et al. 2007) i un monotrema (l'ornitorinc, *Ornithorhynchus anatinus*; Warren et al. 2008) (figura 8).

2. Objectius

En aquesta tesi ens hem proposat els següents objectius:

1. Identificar i anotar el repertori de gens que codifiquen la via de transducció de senyal de la insulina/TOR en els genomes de 12 espècies del gènere *Drosophila* i espècies de 6 vertebrats.

2. Caracteritzar les forces selectives (selecció purificadora i adaptativa) que han actuat durant l'evolució dels gens de la via, a partir de la relació entre els nivells de divergència no sinònima i sinònima ($\omega = d_N/d_S$).

3. Determinar els llinatges de *Drosophila* on s'han produït els esdeveniments de duplicació, pèrdua o pseudogenització de gens.

4. Determinar la possible relació entre els patrons d'evolució molecular dels gens de la via de la insulina/TOR i la posició que ocupen els seus productes a la via.

    4.1. Determinar si existeix una correlació entre la intensitat de la selecció purificadora i la posició dels gens al llarg de l'eix *upstream*/*downstream* de la via.

    4.2. Determinar si els gens que codifiquen proteïnes que interactuen físicament evolucionen sota nivells semblants de selecció purificadora.

    4.3. A *Drosophila*, determinar si els gens que codifiquen proteïnes que interactuen físicament tendeixen a duplicar-se als mateixos llinatges.

    4.4. Determinar si la selecció positiva ha actuat de manera diferencial a parts específiques de la via.

3. Articles

3.1. Article 1
Network-level molecular evolutionary analysis of the insulin/TOR signal transduction pathway across 12 *Drosophila* genomes

David Alvarez-Ponce, Montserrat Aguadé i Julio Rozas

Genome Research 19:234-242 (Febrer 2009)

3.1.1. Resum

La funció biològica està basada en xarxes complexes formades per una gran quantitat de molècules que interactuen entre si. Les propietats evolutives d'aquestes xarxes moleculars i, en particular, l'impacte de la seva arquitectura sobre l'evolució a nivell de seqüència dels seus components individuals són, no obstant això, encara poc compreses. En aquest treball, hem realitzat una anàlisi evolutiva acurada a nivell de xarxa de la via de la insulina/TOR a 12 espècies de *Drosophila*. Hem trobat que els components de la via de la insulina/TOR estan completament conservats en aquestes espècies i que dos gens que actuen a punts de ramificació principals de la via presenten evidències d'haver evolucionat sota selecció positiva. De manera destacable, hem detectat un gradient en la intensitat de la selecció natural al llarg de la via, que augmenta des dels gens que actuen a la part superior de la via fins als que actuen a la part final. També hem trobat que les proteïnes que interactuen físicament tendeixen a evolucionar sota nivells de limitació funcional similars, tot i que aquesta tendència podria ser el resultat de la correlació entre els nivells de limitació funcional i la posició dels gens a la via. Aquests resultats indiquen clarament que els nivells de limitació funcional depenen de la posició que les proteïnes ocupen a la via i, per tant, l'arquitectura de la via limita l'evolució dels gens a nivell de seqüència.

Letter

Network-level molecular evolutionary analysis of the insulin/TOR signal transduction pathway across 12 *Drosophila* genomes

David Alvarez-Ponce, Montserrat Aguadé, and Julio Rozas[1]

Departament de Genètica, Facultat de Biologia, Universitat de Barcelona, Barcelona 08028, Spain

Biological function is based on complex networks consisting of large numbers of interacting molecules. The evolutionary properties of molecular networks and, in particular, the impact of network architecture on the sequence evolution of its individual components are, nonetheless, still poorly understood. Here, we conducted a fine-scale network-level molecular evolutionary analysis of the insulin/TOR pathway across 12 species of *Drosophila*. We found that the insulin/TOR pathway components are completely conserved across these species and that two genes located at major network branch points show evidence for positive selection. Remarkably, we detected a gradient in the strength of purifying selection along the pathway, increasing from the upstream to the downstream genes. We also found that physically interacting proteins tend to have more similar levels of selective constraint, even though this feature might represent a byproduct of the correlation between selective constraint and the pathway position. Our results clearly indicate that the levels of functional constraint do depend on the position of the proteins in the pathway and, consequently, the architecture of the pathway constrains gene sequence evolution.

[Supplemental material is available online at www.genome.org.]

Biological function is based on complex networks consisting of large numbers of molecules. Indeed, genes do not act in isolation but interact in molecular pathways. The evolutionary dynamics of biochemical networks is, moreover, a fundamental issue in systems biology. Establishing the patterns of genetic variation across networks and the impact of natural selection on such variability can provide important insights into the evolutionary forces acting in network evolution. Most evolutionary studies, however, have focused on individual genes or gene families; consequently, the properties and mechanisms underlying network evolution remain largely unknown.

A central question in biological network evolution concerns the role of topology in the evolution of individual network components and, in particular, the effect of the position of an element in the network on the strength of positive and purifying selection. Whole-genome analysis has shown that better connected network elements (e.g., hubs) tend to be more functionally constrained (Fraser et al. 2002; Hahn and Kern 2005; Lemos et al. 2005; Vitkup et al. 2006) and that physically interacting elements tend to exhibit similar levels of selective constraint (Fraser et al. 2002; Lemos et al. 2005). The position of an element in a network, therefore, clearly affects its evolutionary fate. Nevertheless, little research has addressed this question on well-characterized molecular pathways, showing that elements located at network branch points tend to evolve adaptively (Eanes 1999; Flowers et al. 2007). Moreover, the upstream elements in some biochemical pathways are more constrained than those in downstream positions (Rausher et al. 1999; Lu and Rausher 2003; Riley et al. 2003). This kind of selective constraint gradient along the upstream/downstream axis has been explained by the hierarchical organization of these pathways; namely, mutations in upstream genes would generate greater pleiotropic effects than those in genes at the downstream part of the pathway, being therefore more likely to have a deleterious effect.

Biochemical pathways can be classified into three categories: metabolic; transcriptional regulatory; and signal transduction (or signaling) pathways. Signaling pathways transduce signals (such as hormones acting as ligands of extracellular receptors) from outside to inside the cell. The ligand–receptor interaction triggers a cascade of biochemical reactions (often through protein phosphorylation and dephosphorylation). The transduced signal ultimately activates the effector elements of the pathway, which are responsible for mediating the response.

The insulin/TOR (IT) signal transduction pathway plays a central role in many critical biological processes in animals, including organism growth, anabolic metabolism, cell survival, fertility, and lifespan determination (Goberdhan and Wilson 2003; Oldham and Hafen 2003). Both the network topology and the molecular functions of its components have been well characterized in different organisms, including *Drosophila melanogaster* (Supplemental Fig. S1), and are highly conserved across metazoans.

Current knowledge of IT signaling in *D. melanogaster*, with the recent addition of the complete genome sequences for 12 species of the same genus, offers the possibility of conducting a fine-scale evolutionary analysis of a signal transduction pathway. Here, we have studied the molecular evolution of the IT signaling pathway genes of 12 *Drosophila* species within a network-level framework.

Results

Identification of insulin/TOR pathway genes in *Drosophila* genomes

We identified a total of 315 putative orthologs of the 27 *D. melanogaster* IT signaling pathway genes (Table 1) in 11 *Drosophila* genomes. Therefore, we analyzed 342 DNA sequences (Supplemental Table S2). Since current genomic projects include many

[1]Corresponding author.
E-mail jrozas@ub.edu; fax 34-93-4034420.
Article published online before print. Article and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.084038.108.

19:234–242 © 2009 by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/09; www.genome.org

Table 1. Summary statistics used in the multivariate analysis

Gene	Position	Protein length[a]	Percent of analyzed codons[b]	d_N[c]	d_S[c]	ω	Connectivity[d]	Effective no. of codons	mRNA abundance[e]
Akt1	5	530	94.15	0.042	1.059	0.040	4	53.71	144
chico	1	968	89.46	0.159	1.826	0.087	1	56.25	138
dm	7	717	57.04	0.221	2.833	0.078	14	45.77	221
eIF2B-ϵ	7	669	96.11	0.096	1.909	0.050	1	42.99	195
eIF-4E[f]	10	259	85.71	0.035	1.240	0.028	11	45.63	1002
eIF4E-3[f]	—	244	100.00	0.236	1.423	0.166	6	48.33	225
eIF4E-4[f]	—	229	100.00	0.081	1.007	0.080	0	46.74	114
eIF4E-5[f]	—	232	81.47	0.119	1.364	0.087	10	47.21	248
eIF4E-6[f]	—	173	0.00	—	—	—	0	54.24	8
eIF4E-7[f]	—	429	47.79	0.243	2.295	0.106	8	54.46	53
4EHP[f]	—	223	99.55	0.045	0.531	0.085	2	44.93	70
foxo	6	613	89.23	0.041	0.909	0.046	2	43.95	91
gig	6	1847	97.13	0.065	1.780	0.036	0	49.21	93
melt	4	488	96.88	0.036	1.499	0.024	0	47.75	17
Pi3K21B	2	992	91.90	0.142	2.483	0.057	12	48.77	173
Pi3K92E	3	506	95.86	0.102	2.102	0.049	1	46.55	221
Pk61C	4	1088	77.83	0.064	1.397	0.046	8	50.12	276
Pten	—	836	98.25	0.139	0.634	0.220	2	54.36	174
Rheb	7	514	100.00	0.049	2.095	0.024	0	46.42	383
RpS6	10	182	98.01	0.023	0.956	0.024	8	33.48	3186
S6k	9	251	98.16	0.010	0.769	0.013	1	51.81	151
sgg	6	490	71.88	0.035	0.872	0.040	1	48.91	181
step	—	1067	96.72	0.088	1.255	0.070	11	52.64	204
Thor	9	117	100.00	0.034	2.301	0.015	3	39.47	1317
Tor	8	2470	89.12	0.052	2.110	0.025	0	52.77	136
Tsc1	5	1100	93.27	0.086	1.831	0.047	9	48.35	169
CG6904	7	709	100.00	0.014	1.495	0.009	13	44.25	997

[a]Number of amino acids in the *D. melanogaster* protein.
[b]Percentage of the *D. melanogaster* codons used for the ω estimations (the rest represent positions poorly alignable or with alignment gaps).
[c]The d_N and d_S values correspond to the sums across all branches of the *melanogaster* group phylogeny.
[d]Number of PPIs involving each *D. melanogaster* protein.
[e]mRNA signal level in *D. melanogaster* adults (Chintapalli et al. 2007).
[f]Paralogous genes encoding the eukaryotic initiation factor 4E (eIF4E).

unsequenced regions, this should be considered as the minimum number of actual genes. Additionally, recent gene duplication events are difficult to identify given the low divergence between the resulting paralogous copies, which might have been treated as a single copy during genome assembly. Some of the identified sequences are incomplete (they are located in partially sequenced regions), and seven of them reveal some pseudogenization footprint (frameshifts, premature stop codons, or indels; Supplemental Table S2).

All the IT pathway genes studied have orthologs in all 12 genomes, except *eIF4E-6*, which is present only in the *melanogaster* subgroup of *Drosophila*. The *D. melanogaster eIF4E-6* and *4EHP* genes, which belong to a seven-member paralogous group (Table 1), may be either nonfunctional or negative IT signaling regulators (Hernandez et al. 2005). Current results, therefore, suggest that the IT signaling pathway is well conserved across available *Drosophila* genomes. Seventeen IT pathway genes have a 1:1 orthology relationship, while the remaining 10 genes underwent a number of duplication and/or loss events (20 duplications, 1 loss, and 5 pseudogenization events; Fig. 1).

Synonymous and nonsynonymous divergence along the IT pathway

We inferred the impact of natural selection on the IT pathway genes of the *D. melanogaster* group from the nonsynonymous (d_N) to synonymous (d_S) substitution rate ratio ($\omega = d_N/d_S$). The values of ω range from 0.009 for *CG6904* to 0.220 for *Pten* (Table 1). We detected the footprint of positive selection in the *eIF2B-ε*, *Akt1*, and *Tor* genes by comparing the M7 and M8 models (the M7 model assumes a discrete beta distribution for ω [$0 \leq \omega \leq 1$], whereas the M8 model adds an extra class of sites [$\omega > 1$]; Supplemental Table S3). The test is only significant for *eIF2B-ε* and *Akt1* at a false discovery rate (FDR) of 5%.

To study the relationship between the ω values and the architecture of the IT signaling pathway, we evaluated whether: (1)

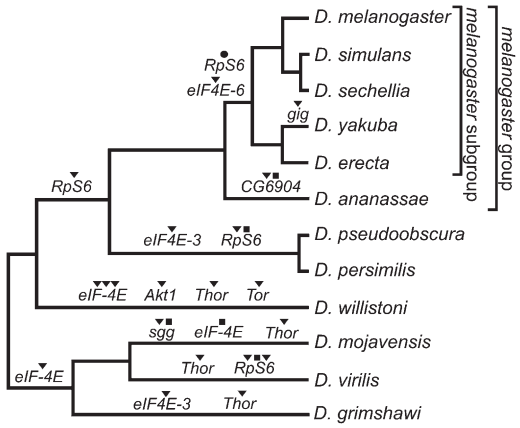

Figure 1. Gene duplication (▼), loss (●), and pseudogenization (■) events detected in the IT pathway across the *Drosophila* phylogeny.

physically interacting elements within the IT pathway have more similar ω values, and (2) the ω values are affected by the position of the elements in the pathway. The first analysis revealed that physically interacting IT pathway proteins (Fig. 2C) tend to evolve at more similar rates: The average absolute difference between the ω values of the physically interacting elements in the IT pathway ($X_\omega = 0.015$) is significantly lower than expected from a network with the same elements and the same number of interactions assigned at random ($\bar{X}_\omega = 0.023$, $P = 0.010$). To establish which ω component is the main contributor to this trend, we conducted the analysis for d_N and d_S independently. The results of the Monte Carlo test showed that the nonsynonymous changes are the main contributors to the tendency ($X_N = 0.031$, $P = 0.004$; $X_S = 0.591$, $P = 0.164$).

We found a significant negative correlation between ω estimates for IT pathway genes and their position in the pathway (computed as the number of steps required to transduce the signal from InR to the other elements; Fig. 2) (Spearman's rank correlation coefficient, $\rho = -0.607$; $P = 0.006$; Fig. 3A). This result suggests that the topology of the IT pathway influences the distribution of selective constraint along it. More specifically, the downstream elements (Fig. 2) have higher levels of selective constraint than the upstream elements. When this analysis was conducted separately for d_N and d_S, we again found that nonsynonymous changes are the main contributors to the tendency (d_N: $\rho = -0.622$, $P = 0.004$, Fig. 3B; d_S: $\rho = -0.165$, $P = 0.499$).

We considered whether the correlation between ω and pathway position was a general trend in the phylogeny or—on the contrary—whether it might be attributable to some specific lineage. To establish this, we analyzed each of the nine lineages (the six external and the three internal branches of the *melanogaster* group phylogeny) separately using the ω values estimated under the free-ratio model (FR). This test is only significant for the *D. yakuba* ($\rho = -0.524$, $P = 0.021$), *D. erecta* ($\rho = -0.511$, $P = 0.025$), and *D. ananassae* ($\rho = -0.729$, $P = 0.0004$) lineages. Even though this correlation is not significant in the six remaining lineages, the ρ statistic is also negative in five of them. We also applied a specific two-ratio branch model to estimate the ω ratios in two groups: one including the *D. yakuba*, *D. erecta*, and *D. ananassae* lineages, and the other comprised of the six remaining lineages. The correlation is significant in the two groups ($\rho = -0.669$, $P = 0.002$; $\rho = -0.455$, $P = 0.050$; respectively), indicating that the negative correlation between the ω values and the position of the elements in the pathway is a phylogeny-wide trend and not caused by any lineage-specific pattern.

The estimates of ω used in the previous analyses were obtained from nucleotide sequence data clearly alignable across the six species of the *melanogaster* group. Since removing the most divergent regions might bias the results, we reanalyzed the data using the noncurated data set (the direct output of the ProbCons alignment software). This analysis does not change the main conclusion, namely, that ω correlates negatively with the position

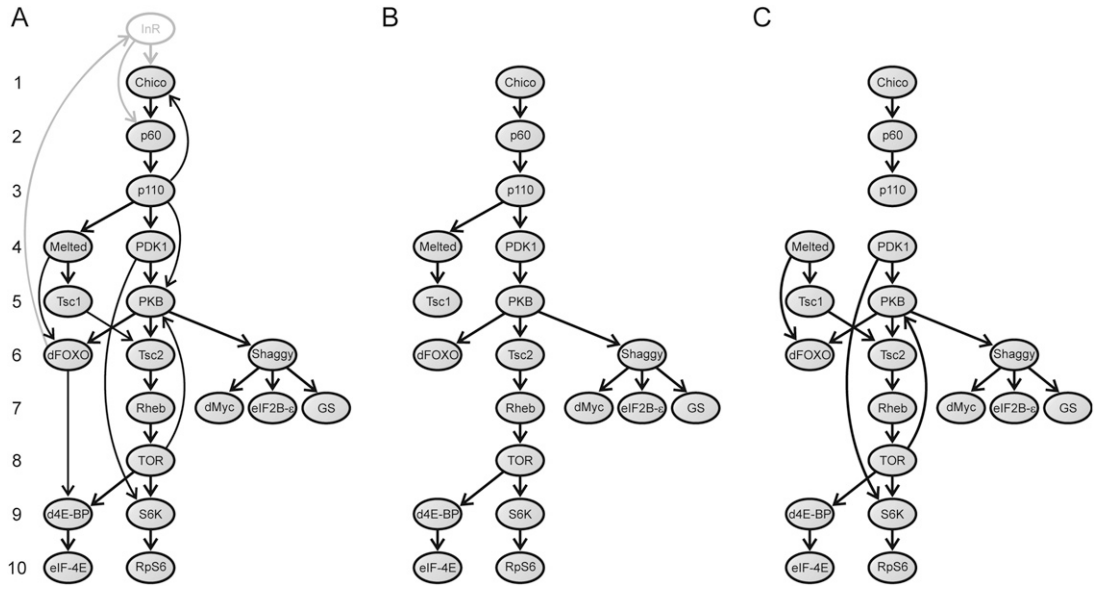

Figure 2. Graphs used in the network-level analysis. (*A*) Directed graph (*G* graph) representing the interactions across the *D. melanogaster* IT pathway elements. Arrows (arcs) indicate the direction of signal transduction. Numbers on the *left* represent the position of the elements in the pathway. (*B*) Graph *T* is a directed spanning tree of *G* used to compute the position of each element in the IT pathway (i.e., the number of signal transduction steps required to transduce the signal from InR to the downstream elements of the pathway). This graph was obtained by removing some arcs from *G* (according to specific biochemical criteria). We eliminated the three arcs involving feedback loops (activation of Chico by PIP_3, which is synthesized by p110; activation of *InR* by the transcription factor dFOXO; phosphorylation of PKB by TOR). Furthermore, if a particular node is reached by different paths (d4E-BP, dFOXO, PKB, S6K, and Tsc2) we considered only one of them. For dFOXO, PKB, and S6K, we chose the longest path, since each of the paths allows the transduction of one necessary but not sufficient signal for the activation/inhibition of these proteins (i.e., the elements need to receive all the signals for activation/inhibition). Indeed, the recruitment of dFOXO to the cell membrane by Melt is a prior step to the phosphorylation (and consequent inhibition) of dFOXO by PKB (the *Akt1* product). In the same way, the recruitment of PKB to the cell membrane through its interaction with PIP_3 (synthesized by p110) is also a prior step to the phosphorylation of PKB by PDK1 (the *Pk61C* product). S6K needs to be phosphorylated by both PDK1 and TOR for full activation (Chou and Blenis 1995; Dufner and Thomas 1999; Avruch et al. 2001). d4E-BP (the *Thor* encoded protein) is an inhibitor of the pathway activated by its transcription factor dFOXO and inhibited by the TOR kinase. Given that only the second interaction activates the pathway, we eliminated the first from the analysis. (*C*) Graph *S* is a subgraph of *G* that includes only the direct physical PPIs between the elements of the IT pathway.

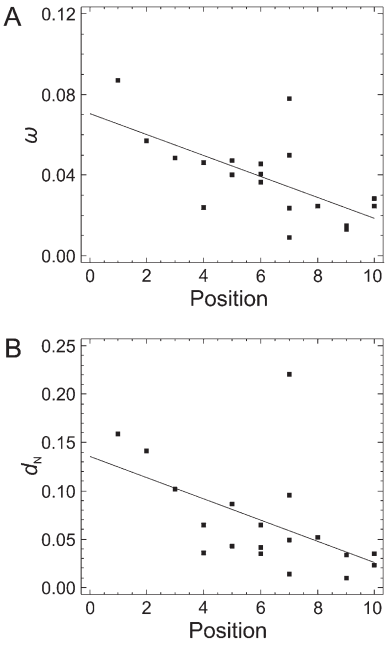

Figure 3. Correlation between the position of the elements in the IT pathway and the ω (*A*) and d_N (*B*) estimates. Continuous lines represent regression lines.

of the elements in the pathway (ρ = –0.559, P = 0.013). Another putative source of bias is the use of an inadequate codon frequency model (the ω values reported here were estimated using the F3×4 codon frequency model; Goldman and Yang 1994). However, the correlation was significant independently of the codon frequency model used to estimate ω (Fequal, F1×4, or F61).

Finally, as the selective constraint of a given gene is known to correlate with different factors, including gene expression level, codon bias, protein length, and connectivity (number of protein–protein interactions [PPIs]), we considered whether these factors could account for the correlation between ω and the position of the elements in the pathway. We found that (1) expression level, codon bias, and protein length show a significant correlation with the position of the elements in the pathway (ρ = 0.484, P = 0.036 for expression level; ρ = –0.497, P = 0.030 for codon bias, measured as the effective number of codons [ENC]; ρ = –0.480, P = 0.037 for protein length; Supplemental Fig. S2B–D), whereas connectivity does not (ρ = 0.083, P = 0.734; Supplemental Fig. S2A), and (2) these factors do not correlate with ω (ρ = –0.213, P = 0.381 for expression level; ρ = 0.207, P = 0.395 for ENC; ρ = 0.354, P = 0.137 for protein length; ρ = 0.213, P = 0.380 for connectivity; Supplemental Fig. S2E–H). Since expression level, codon bias, and protein length are intercorrelated, some of the observed correlations might actually result from indirect rather than from direct effects. We used path analysis to better characterize the relationships among these factors, connectivity, d_N, ω, and the position in the pathway. This joint analysis (Fig. 4) shows that (1) the d_N values are clearly affected by the position of the elements in the pathway (standardized path coefficient, β = –0.481; P = 0.035), even after removing the effects of putatively relevant factors (gene expression level, codon bias, and protein length); (2) connectivity and d_N are positively associated after factoring out the effects of all other variables (β = 0.389, P = 0.027); and (3) apart from d_N, only the gene expression level is significantly influenced by the pathway position (β = 0.484; P = 0.006). The multiple regression model explains 44.4% of the d_N variability. Path analysis using two other causal models (considering gene expression and protein length as exogenous and endogenous variables, respectively) yielded similar results.

Discussion

Distribution of IT pathway genes across *Drosophila* genomes

Our analysis shows that the IT pathway genes underwent 20 gene duplications, one loss, and five pseudogenization events throughout the evolution of the 12 *Drosophila* species (Fig. 1). Nevertheless, all the IT pathway genes have representatives in the 12 *Drosophila* species; the only exception is the *eIF4E-6* gene, which may be a nonfunctional paralog of the *eIF4E* multigene family (Hernandez et al. 2005). The existence of nearly all the genes in all the surveyed species, together with the relatively high selective constraint levels (ω < 0.25), suggests that the IT pathway is functional across all these species.

It has been suggested that proteins that interact with each other tend to show similar phylogenetic patterns of gene duplication and loss, owing to coordinated evolution (Fryxell 1996). Noticeably, we found that some genes encoding physically interacting proteins underwent gene duplication in the same lineages (*Akt1*, *Tor*, *Thor*, and *eIF-4E* in the *D. willistoni* lineage; *eIF4E-3* and *Thor* in the *D. grimshawi* lineage) (Fig. 1). Nevertheless, the null hypothesis of random accumulation of gene duplications across the branches of the phylogeny could not be rejected (Monte Carlo simulation test; P = 0.190).

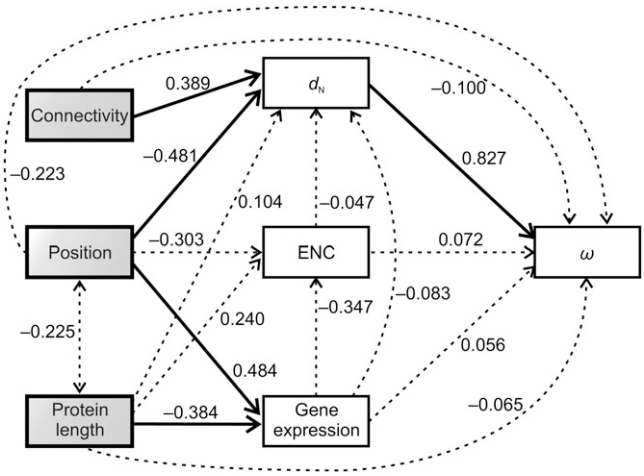

Figure 4. Path analysis used to characterize the relationships among element positions in the IT pathway, nonsynonymous divergence (d_N), d_N/d_S ratio (ω), gene expression level, codon bias (measured by the ENC), protein length, and connectivity. Pathway position, protein length, and connectivity were treated as exogenous variables (those with no explicit causes in the model), while the rest were treated as endogenous variables (those caused by one or more variables in the model). The causal dependencies between variables assumed in the model are represented by single-headed arrows. Correlations between exogenous variables are represented by double-headed arrows. The numbers on the arrows represent the standardized path coefficients (β). Solid and broken lines represent significant and nonsignificant relationships, respectively.

Impact of positive selection

We found that *eIF2B-ϵ*, *Akt1*, and *Tor* genes show the footprint of positive selection (only *eIF2B-ϵ* and *Akt1* after controlling for the FDR). It has been suggested that elements located at branch points of metabolic pathways exert a greater flux control and, therefore, may tend to evolve under positive selection (Eanes 1999; Flowers et al. 2007). If this is so, it should also be true for signal transduction pathways. Interestingly, both PKB and TOR (the encoded products of *Akt1* and *Tor*, respectively) locate at major network branch points (Fig. 2). Upon activation by insulin, p110 catalyzes the synthesis of the membrane lipid PIP_3, which acts as a docking site for a number of pleckstrin homology domain-containing proteins, including PKB. Consistent with the flux control hypothesis, the *Akt1* codons identified as evolving under positive selection are located in the pleckstrin homology domain. Furthermore, since TOR phosphorylates multiple IT pathway elements, it also locates at a major branch point of the IT pathway.

Selective constraints along the IT pathway

We found that physically interacting elements of the IT pathway tend to have more similar ω and d_N values ($P < 0.010$). This pattern, already observed in interactomic-level analyses, has been attributed to the coevolution of amino acids involved in protein interactions (Fraser et al. 2002; Lemos et al. 2005). In our study, however, this pattern might be a byproduct of the current correlation between selective constraint and the pathway position. In fact, after factoring out this effect, the association between ω (and d_N) values of physically interacting elements is no longer significant ($X_\omega = 0.013$, $P = 0.105$; $X_N = 0.030$, $P = 0.057$), although close to the critical value.

Remarkably, our study reveals a robust positive correlation between the position of the elements in the pathway and functional constraint levels. Although both ω and d_N estimates exhibit a statistically significant correlation with the pathway position ($P < 0.006$), results of the path analysis (Fig. 4) clearly indicate that nonsynonymous divergence (d_N) would be the main responsible. A number of factors might underlie the detected correlation between selective constraints and pathway position. First, it has been suggested that regulatory genes tend to evolve faster than structural genes (Tucker and Lundrigan 1993; Whitfield et al. 1993; Purugganan and Wessler 1994; Gaut and Doebley 1997; Rausher et al. 1999), and the structural genes (*eIF-4E*, *RpS6*, *eIF2B-ϵ*, and *CG6904*) in the IT pathway are located downstream. Thus, the observed correlation might be a byproduct of this downstream location of the structural genes. However, the correlation between the position of the elements in the pathway and selective constraint remains significant even after removing these genes from the analysis ($\rho = -0.691$, $P = 0.004$ for ω; $\rho = -0.594$, $P = 0.034$ for d_N). Second, four IT pathway genes (*chico*, *melt*, *Pk61C*, and *Akt1*) that encode proteins with a pleckstrin homology domain are located in the upstream part of the pathway; therefore, relaxed purifying selection in this domain might explain the observed correlation along the pathway. However, the elimination of these genes from the analysis does not affect the results ($\rho = -0.620$, $P = 0.014$ for ω; $\rho = -0.652$, $P = 0.008$ for d_N). Finally, throughout our study we consider that the TOR pathway locates downstream of the insulin pathway. Some experimental studies have questioned this and place some elements of the TOR pathway (Tsc1, Tsc2, Rheb, and TOR) on a route parallel to the insulin pathway (Oldham et al. 2000; Gao et al. 2002; Radimerski et al. 2002; Dong and Pan 2004). Again, the observed correlation remains significant after removing these four elements from the analysis ($\rho = -0.581$, $P = 0.023$ for ω; $\rho = -0.683$, $P = 0.005$ for d_N).

Thus, our results suggest that the structure of the IT pathway constrains the sequence evolution of its components. However, it is not clear what the biological explanation is for the polarity in the strength of purifying selection along the pathway. Diverse factors might affect selective constraints in molecular pathways. For instance, interactomic-level analyses have revealed a negative correlation between evolutionary rate and connectivity (Fraser et al. 2002; Hahn and Kern 2005; Lemos et al. 2005). In contrast, our path analysis uncovered a positive association between d_N and connectivity. Hence, a polarity in the element's connectivity along the pathway might explain the correlation between selective constraint and the pathway position. However, no significant correlation was detected between connectivity (Table 1) and pathway position (Supplemental Fig. S2A); therefore, the connectivity pattern would not explain the correlation between selective constraints and the position of the elements in the pathway. Results based on interactomic data, however, should be taken with caution since current *D. melanogaster* interactomic data is incomplete and unreliable.

Gene expression level, expression breadth (the number of different tissues in which a gene is expressed), codon usage bias, and the length of the encoded proteins can also affect selective constraints. In fact, genes with higher expression levels, higher codon bias, or shorter encoded proteins tend to be more constrained (Duret and Mouchiroud 1999; Pal et al. 2001; Rocha and Danchin 2004; Subramanian and Kumar 2004; Wright et al. 2004; Lemos et al. 2005; Drummond et al. 2006; Ingvarsson 2007). As all IT pathway genes seem to be expressed in all body tissues and structures (Chintapalli et al. 2007), expression breadth cannot account for the pathway polarity of the ω values. A putative higher translation rate of downstream IT pathway genes might justify the observed correlation between ω and the position of the elements in the pathway. In fact, given the signal-amplifying kinetic behavior of the insulin pathway—at least in mammals (Sedaghat et al. 2002), a higher protein abundance is expected in downstream IT pathway elements. On the other hand, shorter protein lengths at the downstream IT pathway part might also generate the observed selective constraint polarity. Interestingly, we detected (1) a positive correlation between the position of the elements in the pathway and both expression level and codon bias (Supplemental Fig. S2B,C) and (2) a negative correlation between protein length and the position of the elements in the pathway (Supplemental Fig. S2D). Namely, downstream IT pathway genes encode shorter and more actively translated proteins. In this pathway, however, none of these factors correlate with ω or d_N (Supplemental Fig. S2F–H). Consequently, these would not be the main factors responsible for the correlation between ω and the position of the elements in the IT pathway. It is conceivable that some coupled effect emerging from codon bias, expression level, and protein length might generate the selective constraint polarity, even though these factors do not correlate with ω or d_N separately. However, path analysis confirms that the relationship between selective constraint and the position of the elements in the pathway is significant even after factoring out the effects of gene expression level, codon bias, protein length, and connectivity (Fig. 4). Consequently, other biological factors are needed to explain the purifying selection polarity along the IT pathway.

The number of molecular pathways in which a gene is involved may affect its functional constraint levels; for instance, highly pleiotropic genes are expected to be more constrained

(Waxman and Peck 1998). Therefore, the distribution of the strength of purifying selection along the upstream/downstream axis of a pathway may be affected by its particular pattern of interconnections with other pathways. A signal transduction pathway receiving signaling inputs from a number of pathways (i.e., with multiple inputs and a single output) is expected to be more constrained at the downstream part given that the downstream elements would be involved in a greater number of pathways (Fig. 5A). Conversely, a network with a branching topology including multiple outputs along the pathway will exhibit the opposite trend in its selective constraint pattern (Fig. 5B). Hence, the balance between the biological relevance of the signaling inputs and outputs might generate a selective constraint polarity along the pathway.

The correlation between functional constraint levels and the position of the elements in the IT pathway might, therefore, be explained by its information flux pattern; in particular, on the basis of the predominance of inputs over outputs along the pathway (in terms of biological relevance). Indeed, even though the IT pathway connection patterns for *Drosophila* are far from being fully known, it does receive inputs from other pathways (Supplemental Table S4). However, some IT pathway elements also transduce signals to other pathways (i.e., there is not just one single output signal) (Supplemental Table S4). Moreover, the biological impact (in terms of fitness) of the interrelations of the IT pathway with these other routes cannot be easily evaluated; therefore, it is difficult to determine whether the effects of signaling inputs overweigh those of the outputs.

Rausher et al. (1999) have shown that the selective constraint levels in the plant anthocyanin biosynthetic pathway also correlate with the position of the elements in the pathway. However, the correlation has the opposite sense to that observed in the IT pathway (i.e., upstream anthocyanin biosynthetic pathway elements are more constrained than those in the downstream part). In this case, the upstream elements are located above major branch points and are consequently involved in the biosynthesis of a greater number of compounds, whereas the downstream genes only affect anthocyanins biosynthesis. The pathway, therefore, has more outputs than inputs (Fig. 5B). Polarity in the selective constraint along the anthocyanin pathway was explained by the involvement of upstream elements in a greater number of biochemical routes (Rausher et al. 1999).

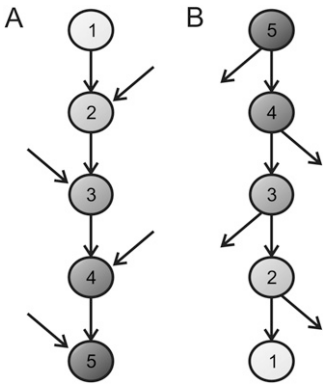

Figure 5. Schematic representation of the selective constraint levels expected along two hypothetical signaling pathways with different connection patterns. (*A*) Pathway receiving multiple signaling inputs along the pathway and with a single output. In this scenario, selective constraint levels will be higher at the downstream part, since the elements are progressively involved in a greater number of pathways. (*B*) Pathway with multiple outputs along the pathway (i.e., with multiple branching points able to transmit information to other pathways). In this scenario, the selective constraint levels will be higher for the upstream elements. The more constrained elements (nodes) are darker. The numbers in the nodes represent the number of pathways in which each element is involved.

The sensitivity of the overall pathway function to the kinetic properties of a given element will also affect selective constraint levels. If genetic variation in the kinetic properties strongly affects the pathway function, the element should be more constrained than if the system works with relative independence from these properties. Therefore, the selective constraint of a protein would be determined not only by its kinetic properties, but also by its position in the pathway and the properties of the interconnected pathway elements. Along these lines, a theoretical analysis conducted in the Ras signaling pathway (Nijhout et al. 2003) predicted that the pathway output would be more sensitive to the upstream enzymes, which therefore should be more constrained. This prediction was supported by DNA polymorphism analysis (Riley et al. 2003). Applying this sensitivity analysis to the IT signaling pathway would probably provide valuable insights into the major biological processes that determine the selective constraints along the pathway.

In summary, even though the biological processes underlying the polarity in the selective constraint levels along the IT pathway remain unclear, our results provide strong evidence that the pathway architecture constrains the molecular evolution of its components. Further work studying the patterns of molecular evolution in pathways encompassing a wide range of topologies and analyzing the biological impact of the interconnection patterns is required to fully understand how network topology constrains the evolution of its components.

Methods

Identification of IT signaling pathway genes in *Drosophila* genomes

The protein coding sequences (CDS) of the IT pathway genes in the *D. melanogaster* genome (release 5.1) (Adams et al. 2000) were retrieved from the FlyBase database (Crosby et al. 2007). Orthologous sequences of these genes in the 11 additional *Drosophila* species with completely sequenced genomes (*D. simulans*, *D. sechellia*, *D. yakuba*, *D. erecta*, *D. ananassae*, *D. pseudoobscura*, *D. persimilis*, *D. willistoni*, *D. mojavensis*, *D. virilis*, and *D. grimshawi*) were obtained from the Assembly, Alignment and Annotation site (http://rana.lbl.gov/drosophila; CAF1 release; Clark et al. 2007). For those genes with multiple splicing isoforms we chose the variant encoding the longest protein among those shared across the 12 species (Supplemental Table S1).

To obtain a bona fide set of genomic orthologous sequences, we curated available preliminary gene annotations and orthologous relationships (GLEAN-R and fuzzy reciprocal BLAST data sets, respectively; Clark et al. 2007). For this purpose, we discarded erroneous automatic orthology assignations; merged those groups of adjacent gene predictions actually corresponding to different regions of a single gene; and annotated coding regions that were unannotated in the original GLEAN-R data set. Putative premature stop codons and frameshift mutations were confirmed by analyzing the genomic trace archives (raw DNA sequence data); these features were discarded if there was at least one sequencing trace without the disrupting mutations. *D. simulans* sequences with incomplete information were curated using DNA sequence data

information from the population genomics project for this species (DPGP Simulans Syntenic Assembly version 2; Begun et al. 2007).

To identify putative unannotated genes, we conducted a two-round search for each orthologous group. First, for each *D. melanogaster* protein we performed a TBLASTN search against all other 11 genomes. Second, each hit (*E*-value $\leq 10^{-5}$) was in silico translated and used as a query for searching the *D. melanogaster* genome. If the best hit in this second round corresponded to the original *D. melanogaster* gene, the sequence was considered an orthologous sequence.

We checked whether identified duplicated genes were artifactual (i.e., attributable to sequencing errors and the consequent erroneous assembly). For this purpose, we used Fisher's exact test to contrast whether the relative number of nucleotide differences between duplicates was similar for silent and nonsynonymous positions. Copies with significantly different ratios were considered to be true paralogs. For the remaining cases, we checked the quality of either the genomic sequences or the trace archives at the mismatch positions, discarding those sequences with poor quality (*phred* score < 20).

We confirmed the orthologous/paralogous relationships of the different *eIF4E* genes in the 12 *Drosophila* species by analyzing the topology of the protein gene tree. Orthologous relationships of highly incomplete sequences were established by colinearity conservation analysis.

Phylogenetic reconstruction

We generated a multiple sequence alignment (MSA) of the amino acid sequences of each orthologous group using the software ProbCons 1.11 (Do et al. 2005). This MSA was used to guide the alignment of the CDS. The resulting CDS alignments were manually improved using the software BioEdit 7.0.5.2. Unreliably aligned regions were removed with Gblocks 0.91b (Castresana 2000) using the default protein alignment parameters.

For each orthologous group, we conducted a bayesian phylogenetic reconstruction using the software MrBayes 3.1.2 (Ronquist and Huelsenbeck 2003), applying the nucleotide substitution model that best fits the data according to the Akaike information criterion. The FindModel program (http://hcv.lanl.gov/content/sequence/findmodel/findmodel.html or http://www.hiv.lanl.gov/content/sequence/findmodel/findmodel.html; an implementation of the MODELTEST software; Posada and Crandall 1998) was used for model selection. When the best-fitting model was the HKY+Γ (not implemented in MrBayes), we used the GTR+Γ model (i.e., the next most complex model implemented in MrBayes). All analyses were conducted allowing for a proportion of sites to be invariable (*I*). The eIF4E protein phylogenetic tree was reconstructed by bayesian inference using the Whelan-Goldman model of amino acid evolution (Whelan and Goldman 2001).

Codon-based analysis

We evaluated the impact of natural selection by estimating nonsynonymous (d_N) and synonymous (d_S) divergence, and their ratio ($\omega = d_N/d_S$) using the program codeml from the PAML 3.15 package (Yang 1997). We restricted this analysis to the six *melanogaster* group species (*D. melanogaster*, *D. simulans*, *D. sechellia*, *D. yakuba*, *D. erecta*, and *D. ananassae*) to avoid saturation at synonymous sites, which could bias the d_S estimates and therefore the ω values, and also because of the impossibility of obtaining reliable alignments for all 12 species. We used MSAs based on 1:1 ortholog sets. In the two cases in which there were more than one gene copy in a given species (i.e., co-orthologs), we used the gene with the most complete sequence or the one without any pseudogenization features (stop codons or frameshift mutations). Only clearly alignable regions of the MSAs were used.

The M0 model (the simplest model, which assumes a single ω value for all lineages and sites) was used for most analyses. We also applied the FR model (which assumes that each lineage has a different ω value) and a specific two-ratio model (assuming two different ω values across the phylogeny). To determine whether some codon positions evolve under positive selection, we compared the M1a and M2a models (Wong et al. 2004) and also the M7 and M8 models (Yang et al. 2000) using the likelihood ratio test (Whelan and Goldman 1999). The FDR associated with multiple testing was controlled at $q = 0.05$ (Benjamini and Hochberg 1995). The Bayes Empirical Bayes approach (Yang et al. 2005) was used to identify the codons evolving under positive selection (posterior probability $\geq$ 95%).

Given the differences between gene trees and the species tree concerning the phylogenetic position of the *D. erecta* and the *D. yakuba* lineages (Pollard et al. 2006), for each orthologous group we used the topology (from the three competing alternatives) that best fits the data according to the M0 model. We conducted all likelihood estimations using three different ω starting values (0.1, 1, and 2) to overcome the problem of multiple local optima. All these analyses were conducted using the F3×4 codon frequency model (Goldman and Yang 1994).

Network-level analysis

We coded the structure of the IT pathway into a directed graph (termed *G*, Fig. 2A) with nodes and arcs representing genes/proteins and signaling (activation/inhibition) interactions, respectively. We restricted the analyses to the intracellular part of the pathway. Elements that do not directly interact with any other element in the graph (PTEN) or which have an unclear position in the pathway (Step; Fuss et al. 2006) were not included in *G*. Additionally, to avoid using redundant information, we considered only one of the seven genes encoding the eIF4E isoforms: the gene with the highest mRNA abundance in *D. melanogaster* (*eIF-4E*; Chintapalli et al. 2007; Hernandez et al. 2005). In total, the resulting *G* graph has 19 nodes connected by 25 arcs. Twenty of these interactions are physical—direct PPIs, four are metabolic (p110 catalyzes the synthesis of the membrane phospholipid PIP_3, which recruits Chico, Melted, PDK1, and PKB proteins to the cell membrane), and the other involves the activation of the *Thor* gene by the dFOXO transcription factor.

We generated two subgraphs of *G* (termed *S* and *T*) by removing some arcs. The *S* graph contains only the 20 physical PPIs (Fig. 2C) and was used to contrast whether levels of selective constraint and patterns of gene duplication are more similar for physically interacting proteins. *T* is a directed spanning tree of *G* obtained according to biochemical criteria; in this graph, Chico is in the root (upstream) while the effectors of the pathway are downstream (Fig. 2B). This graph was used to establish the position of the elements in the pathway, defined as the number of steps required to transduce the signal from InR to the other elements (the maximum number of steps was 10).

To establish whether physically interacting proteins in the IT signaling pathway exhibit similar levels of selective constraint, we applied the Monte Carlo method described in Fraser et al. (2002) to the *S* graph. For the analysis we used the *X* statistic, defined as

$$X = \frac{1}{n}\sum_{i=1}^{n} |x_{i1} - x_{i2}|$$

where x_{i1} and x_{i2} are the evolutionary parameters (either d_N, d_S, or ω; the analysis was conducted separately for the three parameters) of the two genes encoding interacting proteins (1 and 2) at pair *i*,

and n is the total number of interacting protein pairs (20 for the IT pathway). The statistical significance of X was determined by generating 100,000 randomizations of S. Each randomization had the same 19 nodes as S, and the same number of arcs ($n = 20$). Each arc was generated by randomly choosing two distinct nodes from S. To factor out the effect of the correlation between the pathway position and selective constraint, we conducted a modification of this Monte Carlo test. After fitting a linear model to the data (i.e., obtaining the regression equation relating the pathway position and either ω or d_N), we used the residuals of the linear model to obtain the X statistic value (i.e., for each gene we used as evolutionary parameter the difference between the observed and predicted selective constraint—ω or d_N—values).

We carried out an additional Monte Carlo test to determine whether the genes encoding physically interacting proteins tend to duplicate in the same phylogenetic branch. We used as statistic the number of gene pairs encoding physically interacting proteins that duplicated in the same phylogenetic branch. The statistical significance was evaluated on the basis of 100,000 replicates. In each replicate we incorporated 20 duplication events (sampled with replacement from that observed in our data; Fig. 1) across the 22 branches of the phylogenetic tree. Each duplication event was incorporated into a given branch with a probability proportional to its branch length. For the analysis we used the *Drosophila* tree topology and branch lengths reported in Russo et al. (1995).

Multivariate analysis

We performed a multivariate analysis considering d_N, ω, the pathway position, and some parameters influencing purifying selection levels (expression level, codon bias, protein length, and connectivity). First, we evaluated whether these parameters correlated using Spearman's rank correlation coefficient (ρ). Later, we analyzed the data using path analysis, an extension of multiple regression analysis that allows decomposing the regression coefficients into their direct and indirect components by considering an underlying user-defined causal model, and to assess the statistical significance of the relevant direct components. This analysis was conducted using the Amos 6.0 software.

Connectivity was estimated as the number of PPIs involving each *D. melanogaster* IT pathway protein. Putative PPIs dealing with these proteins were obtained from Giot et al. (2003). mRNA abundance in the *D. melanogaster* adult body of each gene was obtained from the FlyAtlas database (Chintapalli et al. 2007). These data were log-transformed for the path analysis to improve normality. The codon usage bias of each orthologous group was measured as the median of ENC (Wright 1990) of the six *melanogaster* group species. ENC values of each sequence were obtained using the DnaSP 4.20.1 software (Rozas et al. 2003).

Acknowledgments

We thank the anonymous reviewers for helpful comments and suggestions. This work was supported by grants BFU2004-02253, BFU2007-62927, and BFU2007-63228 from the Ministerio de Educación y Ciencia (Spain); grant 2005SRG-00166 from the Comissió Interdepartamental de Recerca i Innovació Tecnològica (Spain); and special support (Distinció per la Promoció de la Recerca Universitària, to M.A.) from the Generalitat de Catalunya (Spain). D.A-P. was supported by a predoctoral fellowship from the Ministerio de Educación y Ciencia (Spain).

References

Adams, M.D., Celniker, S.E., Holt, R.A., Evans, C.A., Gocayne, J.D., Amanatides, P.G., Scherer, S.E., Li, P.W., Hoskins, R.A., Galle, R.F., et al. 2000. The genome sequence of *Drosophila melanogaster*. *Science* **287:** 2185–2195.

Avruch, J., Belham, C., Weng, Q., Hara, K., and Yonezawa, K. 2001. The p70 S6 kinase integrates nutrient and growth signals to control translational capacity. *Prog. Mol. Subcell. Biol.* **26:** 115–154.

Begun, D.J., Holloway, A.K., Stevens, K., Hillier, L.W., Poh, Y.P., Hahn, M.W., Nista, P.M., Jones, C.D., Kern, A.D., Dewey, C.N., et al. 2007. Population genomics: Whole-genome analysis of polymorphism and divergence in *Drosophila simulans*. *PLoS Biol.* **5:** e310. doi: 10.1371/journal.pbio.0050310.

Benjamini, Y. and Hochberg, Y. 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Statist. Soc. B* **57:** 289–300.

Castresana, J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol. Biol. Evol.* **17:** 540–552.

Clark, A.G., Eisen, M.B., Smith, D.R., Bergman, C.M., Oliver, B., Markow, T.A., Kaufman, T.C., Kellis, M., Gelbart, W., Iyer, V.N., et al. 2007. Evolution of genes and genomes on the *Drosophila* phylogeny. *Nature* **450:** 203–218.

Crosby, M.A., Goodman, J.L., Strelets, V.B., Zhang, P., and Gelbart, W.M. 2007. FlyBase: Genomes by the dozen. *Nucleic Acids Res.* **35:** D486–D491.

Chintapalli, V.R., Wang, J., and Dow, J.A. 2007. Using FlyAtlas to identify better *Drosophila melanogaster* models of human disease. *Nat. Genet.* **39:** 715–720.

Chou, M.M. and Blenis, J. 1995. The 70 kDa S6 kinase: Regulation of a kinase with multiple roles in mitogenic signalling. *Curr. Opin. Cell Biol.* **7:** 806–814.

Do, C.B., Mahabhashyam, M.S., Brudno, M., and Batzoglou, S. 2005. ProbCons: Probabilistic consistency-based multiple sequence alignment. *Genome Res.* **15:** 330–340.

Dong, J. and Pan, D. 2004. Tsc2 is not a critical target of Akt during normal *Drosophila* development. *Genes & Dev.* **18:** 2479–2484.

Drummond, D.A., Raval, A., and Wilke, C.O. 2006. A single determinant dominates the rate of yeast protein evolution. *Mol. Biol. Evol.* **23:** 327–337.

Dufner, A. and Thomas, G. 1999. Ribosomal S6 kinase signaling and the control of translation. *Exp. Cell Res.* **253:** 100–109.

Duret, L. and Mouchiroud, D. 1999. Expression pattern and, surprisingly, gene length shape codon usage in *Caenorhabditis*, *Drosophila*, and *Arabidopsis*. *Proc. Natl. Acad. Sci.* **96:** 4482–4487.

Eanes, W.F. 1999. Analysis of selection on enzyme polymorphisms. *Rev. Ecol. Syst.* **30:** 301–326.

Flowers, J.M., Sezgin, E., Kumagai, S., Duvernell, D.D., Matzkin, L.M., Schmidt, P.S., and Eanes, W.F. 2007. Adaptive evolution of metabolic pathways in *Drosophila*. *Mol. Biol. Evol.* **24:** 1347–1354.

Fraser, H.B., Hirsh, A.E., Steinmetz, L.M., Scharfe, C., and Feldman, M.W. 2002. Evolutionary rate in the protein interaction network. *Science* **296:** 750–752.

Fryxell, K.J. 1996. The coevolution of gene family trees. *Trends Genet.* **12:** 364–369.

Fuss, B., Becker, T., Zinke, I., and Hoch, M. 2006. The cytohesin Steppke is essential for insulin signalling in *Drosophila*. *Nature* **444:** 945–948.

Gao, X., Zhang, Y., Arrazola, P., Hino, O., Kobayashi, T., Yeung, R.S., Ru, B., and Pan, D. 2002. Tsc tumour suppressor proteins antagonize amino-acid-TOR signalling. *Nat. Cell Biol.* **4:** 699–704.

Gaut, B.S. and Doebley, J.F. 1997. DNA sequence evidence for the segmental allotetraploid origin of maize. *Proc. Natl. Acad. Sci.* **94:** 6809–6814.

Giot, L., Bader, J.S., Brouwer, C., Chaudhuri, A., Kuang, B., Li, Y., Hao, Y.L., Ooi, C.E., Godwin, B., Vitols, E., et al. 2003. A protein interaction map of *Drosophila melanogaster*. *Science* **302:** 1727–1736.

Goberdhan, D.C. and Wilson, C. 2003. The functions of insulin signaling: Size isn't everything, even in *Drosophila*. *Differentiation* **71:** 375–397.

Goldman, N. and Yang, Z. 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol. Biol. Evol.* **11:** 725–736.

Hahn, M.W. and Kern, A.D. 2005. Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks. *Mol. Biol. Evol.* **22:** 803–806.

Hernandez, G., Altmann, M., Sierra, J.M., Urlaub, H., del Corral, R.D., Schwartz, P., and Rivera-Pomar, R. 2005. Functional analysis of seven genes encoding eight translation initiation factor 4E (eIF4E) isoforms in *Drosophila*. *Mech. Dev.* **122:** 529–543.

Ingvarsson, P.K. 2007. Gene expression and protein length influence codon usage and rates of sequence evolution in *Populus tremula*. *Mol. Biol. Evol.* **24:** 836–844.

Lemos, B., Bettencourt, B.R., Meiklejohn, C.D., and Hartl, D.L. 2005. Evolution of proteins and gene expression levels are coupled in *Drosophila* and are independently associated with mRNA abundance, protein length, and number of protein–protein interactions. *Mol. Biol. Evol.* **22:** 1345–1354.

Lu, Y. and Rausher, M.D. 2003. Evolutionary rate variation in anthocyanin pathway genes. *Mol. Biol. Evol.* **20:** 1844–1853.

Nijhout, H.F., Berg, A.M., and Gibson, W.T. 2003. A mechanistic study of evolvability using the mitogen-activated protein kinase cascade. *Evol. Dev.* **5:** 281–294.
Oldham, S. and Hafen, E. 2003. Insulin/IGF and target of rapamycin signaling: A TOR de force in growth control. *Trends Cell Biol.* **13:** 79–85.
Oldham, S., Montagne, J., Radimerski, T., Thomas, G., and Hafen, E. 2000. Genetic and biochemical characterization of dTOR, the *Drosophila* homolog of the target of rapamycin. *Genes & Dev.* **14:** 2689–2694.
Pal, C., Papp, B., and Hurst, L.D. 2001. Highly expressed genes in yeast evolve slowly. *Genetics* **158:** 927–931.
Pollard, D.A., Iyer, V.N., Moses, A.M., and Eisen, M.B. 2006. Widespread discordance of gene trees with species tree in *Drosophila*: Evidence for incomplete lineage sorting. *PLoS Genet.* **2:** e173. doi: 10.1371/journal.pgen.0020173.
Posada, D. and Crandall, K.A. 1998. MODELTEST: Testing the model of DNA substitution. *Bioinformatics* **14:** 817–818.
Purugganan, M.D. and Wessler, S.R. 1994. Molecular evolution of the plant R regulatory gene family. *Genetics* **138:** 849–854.
Radimerski, T., Montagne, J., Rintelen, F., Stocker, H., van der Kaay, J., Downes, C.P., Hafen, E., and Thomas, G. 2002. dS6K-regulated cell growth is dPKB/dPI(3)K-independent, but requires dPDK1. *Nat. Cell Biol.* **4:** 251–255.
Rausher, M.D., Miller, R.E., and Tiffin, P. 1999. Patterns of evolutionary rate variation among genes of the anthocyanin biosynthetic pathway. *Mol. Biol. Evol.* **16:** 266–274.
Riley, R.M., Jin, W., and Gibson, G. 2003. Contrasting selection pressures on components of the Ras-mediated signal transduction pathway in *Drosophila*. *Mol. Ecol.* **12:** 1315–1323.
Rocha, E.P. and Danchin, A. 2004. An analysis of determinants of amino acids substitution rates in bacterial proteins. *Mol. Biol. Evol.* **21:** 108–116.
Ronquist, F. and Huelsenbeck, J.P. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* **19:** 1572–1574.
Rozas, J., Sanchez-DelBarrio, J.C., Messeguer, X., and Rozas, R. 2003. DnaSP, DNA polymorphism analyses by the coalescent and other methods. *Bioinformatics* **19:** 2496–2497.
Russo, C.A., Takezaki, N., and Nei, M. 1995. Molecular phylogeny and divergence times of drosophilid species. *Mol. Biol. Evol.* **12:** 391–404.
Sedaghat, A.R., Sherman, A., and Quon, M.J. 2002. A mathematical model of metabolic insulin signaling pathways. *Am. J. Physiol. Endocrinol. Metab.* **283:** E1084–E1101.
Subramanian, S. and Kumar, S. 2004. Gene expression intensity shapes evolutionary rates of the proteins encoded by the vertebrate genome. *Genetics* **168:** 373–381.
Tucker, P.K. and Lundrigan, B.L. 1993. Rapid evolution of the sex determining locus in Old World mice and rats. *Nature* **364:** 715–717.
Vitkup, D., Kharchenko, P., and Wagner, A. 2006. Influence of metabolic network structure and function on enzyme evolution. *Genome Biol.* **7:** R39. doi: 10.1186/gb-2006-7-5-r39.
Waxman, D. and Peck, J.R. 1998. Pleiotropy and the preservation of perfection. *Science* **279:** 1210–1213.
Whelan, S. and Goldman, N. 1999. Distributions of statistics used for the comparison of models of sequence evolution in phylogenetics. *Mol. Biol. Evol.* **16:** 1292–1299.
Whelan, S. and Goldman, N. 2001. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Mol. Biol. Evol.* **18:** 691–699.
Whitfield, L.S., Lovell-Badge, R., and Goodfellow, P.N. 1993. Rapid sequence evolution of the mammalian sex-determining gene SRY. *Nature* **364:** 713–715.
Wong, W.S., Yang, Z., Goldman, N., and Nielsen, R. 2004. Accuracy and power of statistical methods for detecting adaptive evolution in protein coding sequences and for identifying positively selected sites. *Genetics* **168:** 1041–1051.
Wright, F. 1990. The "effective number of codons" used in a gene. *Gene* **87:** 23–29.
Wright, S.I., Yau, C.B., Looseley, M., and Meyers, B.C. 2004. Effects of gene expression on molecular evolution in *Arabidopsis thaliana* and *Arabidopsis lyrata*. *Mol. Biol. Evol.* **21:** 1719–1726.
Yang, Z. 1997. PAML: A program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci.* **13:** 555–556.
Yang, Z., Nielsen, R., Goldman, N., and Pedersen, A.M. 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics* **155:** 431–449.
Yang, Z., Wong, W.S., and Nielsen, R. 2005. Bayes empirical bayes inference of amino acid sites under positive selection. *Mol. Biol. Evol.* **22:** 1107–1118.

Received July 31, 2008; accepted in revised form November 20, 2008.

Supplementary information

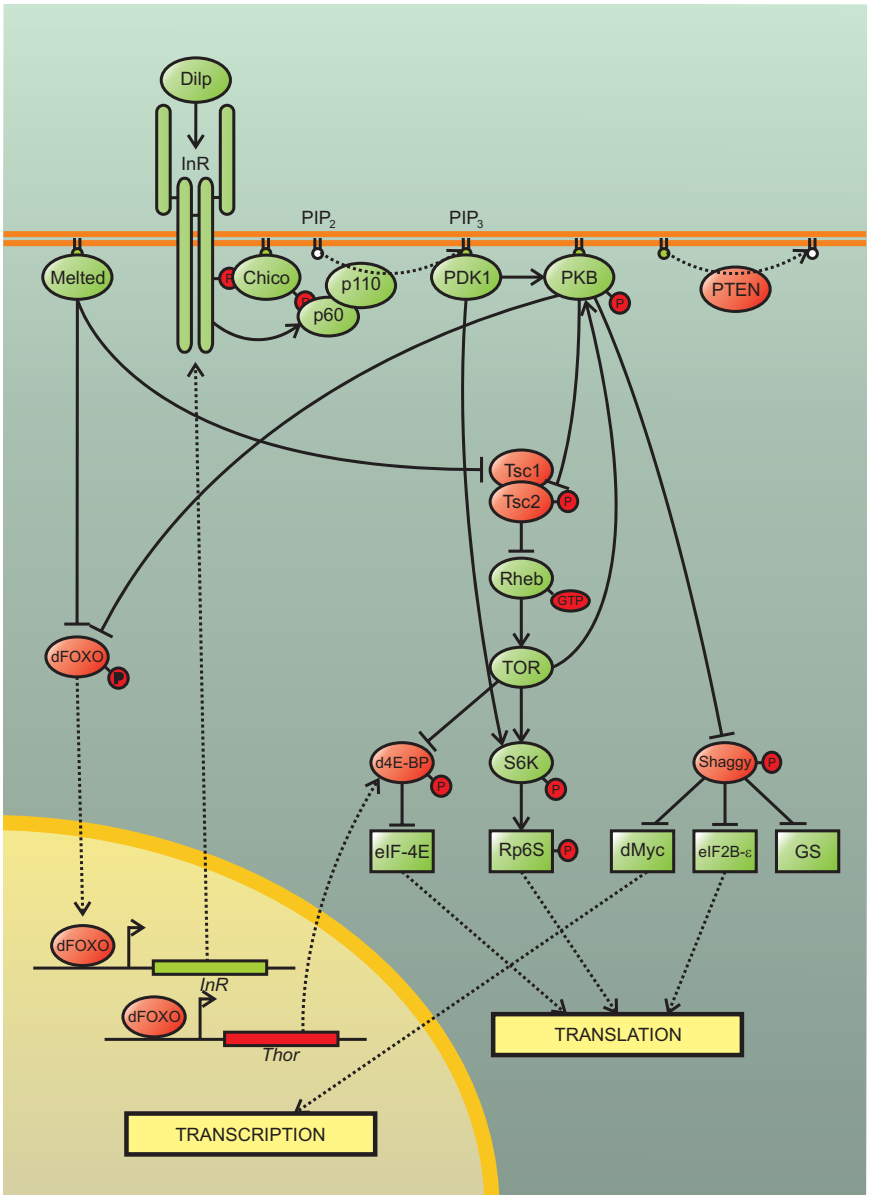

Figure S1. Insulin/TOR pathway in *D. melanogaster*. In the presence of insulin, the insulin receptor (InR) undergoes autophosphorylation providing docking sites for a protein complex; this complex catalyzes the phosphorylation of the membrane lipid phosphatidylinositol-4,5-bisphosphate (PIP_2) to phosphatidylinositol-3,4,5-trisphosphate (PIP_3). The secondary messenger PIP_3 recruits a series of pleckstrin homology domain-containing proteins to the plasma membrane including PKB, which becomes phosphorylated at the plasma membrane. The phosphorylated form of PKB phosphorylates downstream elements, thus unleashing a reactions cascade that activates a series of effectors, including transcription factors and proteins involved in translation and in anabolic metabolism. Activatory and inhibitory elements are represented in green and in red, respectively. Solid lines indicate physical protein-protein interactions (PPIs). Activating and inhibiting interactions are represented by arrows and by lines ending in 'T', respectively. The pathway final effector proteins are represented by rectangles. Recently, Bai *et al.* have shown that, in mammalian cells, Rheb activates mTOR by binding FKBP38 (an inhibitor of mTOR) (Bai et al. 2007). Even though there is a putative FKBP38 ortholog in the *D. melanogaster* genome (*CG5482*), currently it is unknown if this gene is really involved in the IT pathway. Furthermore, previous analysis suggested that the activation of mTOR by Rheb takes place through direct Rheb-TOR interaction (Long et al. 2005). Therefore, we have not considered in the analyses the *CG5482* gene. Including this gene in the analyses, however, does not change the main conclusions of our work.

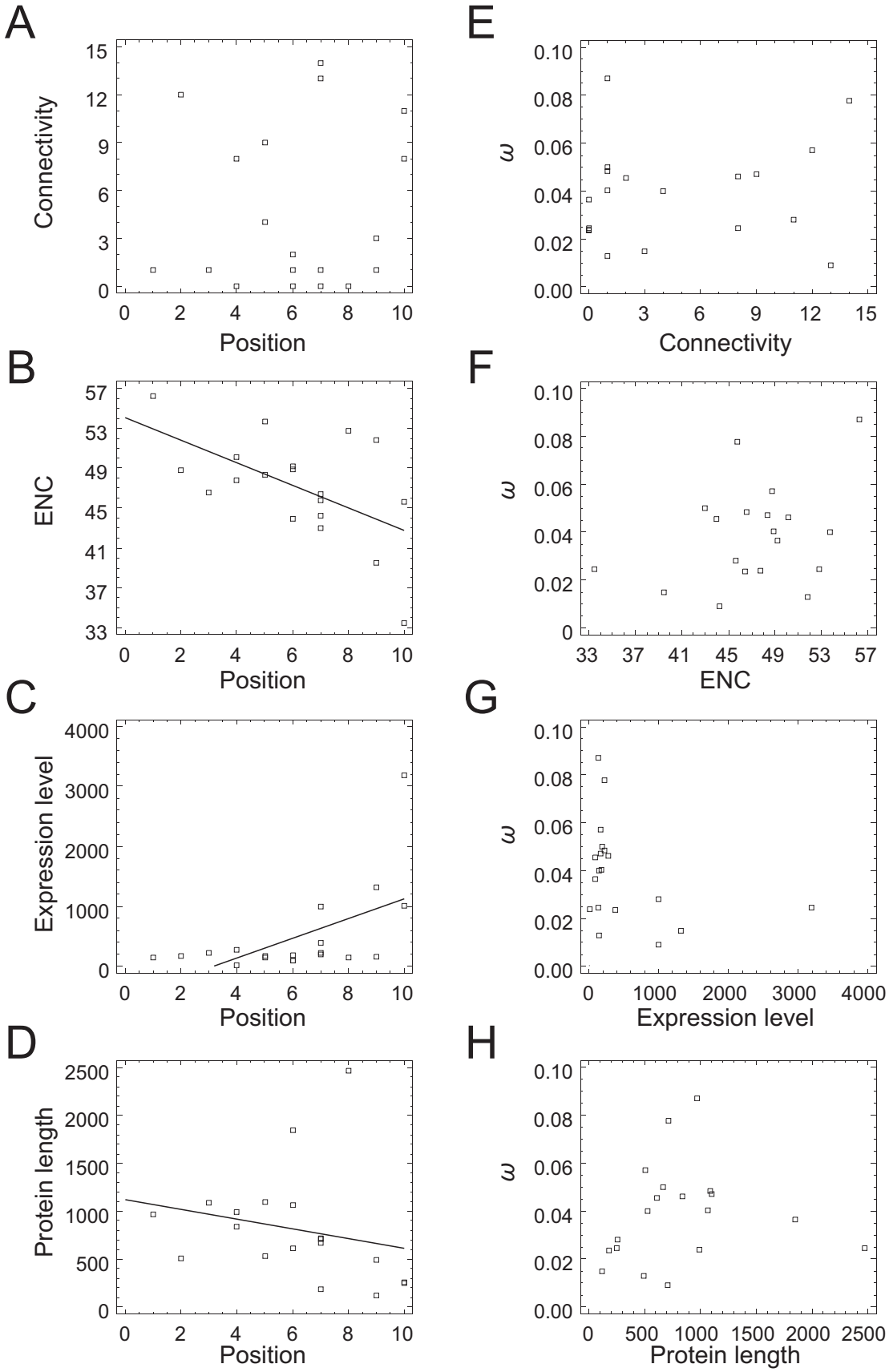

Figure S2. Correlations between different factors (connectivity, codon bias, gene expression levels and protein length) and either the element position in the insulin/TOR pathway (A–D) or the ω values (E–H). Regression lines are represented only for the significant correlations.

Table S1. Genes involved in the *D. melanogaster* insulin/TOR signaling pathway.

Gene	Accession number	Protein	Number of isoforms	Chosen isoform	Protein length	Chromosome	Position
Akt1	CG4006	PKB	2	A	530	3R	11,927,872
chico	CG5686	Chico	1	A	968	2L	10,244,329
dm	CG10798	dMyc	1	A	717	X	3,274,192
eIF2B-ε	CG3806	eIF2B-ε	1	A	669	X	1,815,720
eIF-4E	CG4035	eIF4E-1,2	2	A	259	3L	9,393,583
eIF4E-3	CG8023	eIF4E-3	1	A	244	3L	8,223,001
eIF4E-4	CG10124	eIF4E-4	1	A	229	3L	6,658,069
eIF4E-5	CG8277	eIF4E-5	1	A	232	3L	7,888,543
eIF4E-6	CG1442	eIF4E-6	1	A	173	3R	24,852,462
eIF4E-7	CG32859	eIF4E-7	1	A	429	X	1,053,493
4EHP	CG33100	eIF4E-8	1	A	223	3R	19,911,879
foxo	CG3143	dFOXO	2	B	613	3R	9,899,992
gig	CG6975	Tsc2	1	A	1847	3L	20,127,198
melt	CG8624	Melted	2	A	992	3L	7,134,079
Pi3K21B	CG2699	p60	1	A	506	2L	300,531
Pi3K92E	CG4141	p110	1	A	1088	3R	16,457,356
Pk61C	CG1210	PDK1	4	D	836	3L	136,821
Pten	CG5671	PTEN	3	D	514	2L	10,258,031
Rheb	CG1081	Rheb	1	A	182	3R	1,395,392
RpS6	CG10944	RpS6	3	C	251	X	7,794,475
S6k	CG10539	S6k	1	A	490	3L	5,798,515
sgg	CG2621	Shaggy	6	D	1067	X	2,561,010
step	CG11628	Step	2	B	488	2L	21,744,219
Thor	CG8846	d4E-BP	1	A	117	2L	3,479,004
Tor	CG5092	TOR	1	A	2470	2L	13,008,859
Tsc1	CG6147	Tsc1	1	A	1100	3R	19,958,219
CG6904[a]	CG6904	GS[a]	2	A	709	3R	10,969,512

[a]Putative ortholog of the glucogen synthase.

Table S2. Copy number of the insulin/TOR signaling pathway genes in 12 *Drosophila* genomes

Gene	Dmel	Dsim	Dsec	Dyak	Dere	Dana	Dpse	Dper	Dwil	Dmoj	Dvir	Dgri
Akt1	1	1	1	1	1	1	1	1	2[a]	1	1	1
chico	1	1	1	1	1	1	1	1	1	1	1	1
dm	1	1	1	1	1	1	1	1	1	1	1	1
eIF2B-ε	1	1	1	1	1	1	1	1	1	1	1	1
eIF-4E	1	1	1	1	1	1	1	1	4	1+1[b]	2	2
eIF4E-3	1	1	1	1	1	1	2	2	1	1	1	2
eIF4E-4	1	1	1	1	1	1	1	1	1	1	1	1
eIF4E-5	1	1	1	1	1	1	1	1	1	1	1	1
eIF4E-6	1	1	1	1	1	0	0	0	0	0	0	0
eIF4E-7	1	1	1	1	1	1	1	1	1	1	1	1
4EHP	1	1	1	1	1	1	1	1	1	1	1	1
foxo	1	1	1	1	1	1	1	1	1	1	1	1
gig	1	1	1	2	1	1	1	1	1	1	1	1
melt	1	1	1	1	1	1	1	1	1	1	1	1
Pi3K21B	1	1	1	1	1	1	1	1	1	1	1	1
Pi3K92E	1	1	1	1	1	1	1	1	1	1	1	1
Pk61C	1	1	1	1	1	1	1	1	1	1	1	1
Pten	1	1	1	1	1	1	1	1	1	1	1	1
Rheb	1	1	1	1	1	1	1	1	1	1	1	1
RpS6	1	1	1	1	1	2	2+1[b]	2+1[b]	1	1	1+2[b]	1
S6k	1	1	1	1	1	1	1	1	1	1	1	1
sgg	1	1	1	1	1	1	1	1	1	1+1[b]	1	1
step	1	1	1	1	1	1	1	1	1	1	1	1
Thor	1	1	1	1	1	1	1	1	2	2[a]	2[a]	2[a]
Tor	1	1	1	1	1	1	1	1	2[a]	1	1	1
Tsc1	1	1	1	1	1	1	1	1	1	1	1	1
CG6904	1	1	1	1	1	1+1[ab]	1	1	1	1	1	1

Dmel, *D. melanogaster*; Dsim, *D. simulans*; Dsec, *D. sechellia*; Dyak, *D. yakuba*; Dere, *D. erecta*; Dana, *D. ananassae*; Dpse, *D. pseudoobscura*; Dper, *D. persimilis*; Dwil, *D. willistoni*; Dmoj, *D. mojavensis*; Dvir, *D. virilis*; Dgri, *D. grimshawi*.

[a]Copies located in tandem in the same genomic scaffold.

[b]Pseudogene, DNA sequence with a frameshift or covering less than 40% of the *D. melanogaster* ortholog.

Table S3. Phylogenetic analysis by maximum likelihood of the insulin/TOR pathway genes

Gene	ℓ_{M0}	ℓ_{FR}	ℓ_{M1a}	ℓ_{M2a}	ℓ_{M3}	ℓ_{M7}	ℓ_{M8}	$2(\ell_{FR}-\ell_{M0})$		$2(\ell_{M2a}-\ell_{M1a})$	$2(\ell_{M8}-\ell_{M7})$		$2(\ell_{M3}-\ell_{M0})$	
Akt1	–3,233.47	–3,219.85	–3,194.43	–3,191.90	–3,191.72	–3,197.52	–3,191.72	27.25	***	5.06	11.60	**	83.51	***
chico	–7,055.41	–7,046.34	–7,003.48	–7,003.48	–6,991.73	–6,993.51	–6,992.98	18.14	*	0.00	1.06		127.35	***
dm	–3,620.02	–3,611.08	–3,569.89	–3,569.89	–3,564.14	–3,565.58	–3,565.39	17.89	**	0.00	0.39		111.77	***
eIF2B-ε	–4,679.84	–4,673.55	–4,656.33	–4,655.25	–4,654.48	–4,662.72	–4,655.32	12.60		2.16	14.80	***	50.74	***
eIF-4E	–1,368.58	–1,365.22	–1,360.96	–1,360.96	–1,358.49	–1,358.49	–1,358.49	6.73		0.00	0.00		20.19	***
eIF4E-3	–2,059.13	–2,049.09	–2,035.92	–2,035.92	–2,035.00	–2,035.98	–2,035.18	20.08	**	0.00	1.60		48.26	***
eIF4E-4	–1,565.39	–1,561.53	–1,558.78	–1,558.78	–1,555.94	–1,556.55	–1,556.55	7.72		0.00	0.00		18.90	***
eIF4E-5	–1,449.04	–1,427.93	–1,446.09	–1,446.09	–1,444.33	–1,444.42	–1,444.42	42.22	***	0.00	0.00		9.43	
eIF4E-6[a]	–	–	–	–	–	–	–	–		–	–		–	
eIF4E-7	–1,965.06	–1,918.53	–1,951.28	–1,951.28	–1,950.66	–1,953.51	–1,951.06	93.05	***	0.00	4.89		28.79	***
4EHP	–1,311.36	–1,294.92	–1,310.00	–1,310.00	–1,309.86	–1,309.98	–1,309.88	32.88	***	0.00	0.21		2.99	
foxo	–3,429.60	–3,420.59	–3,420.45	–3,420.45	–3,420.21	–3,420.33	–3,420.22	18.02	**	0.00	0.23		18.79	***
gig	–13,110.05	–13,103.01	–13,074.83	–13,074.60	–13,063.34	–13,065.68	–13,063.42	14.08		0.48	4.51		93.42	***
melt	–6,518.80	–6,515.25	–6,496.18	–6,496.18	–6,494.46	–6,496.07	–6,494.49	7.11		0.00	3.17		48.67	***
Pi3K21B	–3,562.99	–3,557.21	–3,538.34	–3,538.34	–3,531.28	–3,531.84	–3,531.62	11.57		0.00	0.44		63.42	***
Pi3K92E	–7,852.60	–7,843.01	–7,822.97	–7,822.97	–7,810.25	–7,810.43	–7,810.43	19.20	**	0.00	0.00		84.71	***
Pk61C	–4,576.43	–4,573.15	–4,567.55	–4,567.55	–4,564.51	–4,564.72	–4,564.57	6.55		0.00	0.30		23.85	***
Pten	–3,599.14	–3,588.06	–3,570.06	–3,569.09	–3,569.09	–3,571.22	–3,569.10	22.16	**	1.94	4.24		60.09	***
Rheb	–1,244.78	–1,239.81	–1,244.20	–1,244.20	–1,242.16	–1,242.25	–1,242.25	9.93		0.00	0.00		5.23	
RpS6	–1,383.46	–1,377.25	–1,374.83	–1,374.40	–1,374.40	–1,376.67	–1,374.43	12.42		0.85	4.48		18.12	***
S6k	–2,769.89	–2,765.16	–2,769.88	–2,769.86	–2,769.86	–2,769.94	–2,769.91	9.48		0.04	0.06		0.07	
sgg	–4,760.04	–4,757.31	–4,728.18	–4,728.18	–4,727.92	–4,728.22	–4,728.00	5.47		0.00	0.44		64.23	***
step	–3,534.06	–3,521.30	–3,516.42	–3,516.42	–3,513.24	–3,513.79	–3,513.56	25.52	***	0.00	0.46		41.65	***
Thor	–799.33	–797.34	–799.33	–799.33	–799.33	–799.49	–799.49	3.96		0.00	0.00		0.00	
Tor	–15,824.18	–15,789.18	–15,754.01	–15,754.01	–15,747.90	–15,751.67	–15,747.39	69.99	***	0.00	8.57	*	152.56	***
Tsc1	–7,365.47	–7,356.64	–7,348.84	–7,348.84	–7,342.41	–7,343.34	–7,342.67	17.66	**	0.00	1.34		46.12	***
CG6904	–4,305.75	–4,297.50	–4,297.89	–4,297.89	–4,296.97	–4,298.61	–4,297.56	16.50	*	0.00	2.09		17.56	***

ℓ_i, log-likelihood of the observed data under the evolutionary model *i*. *, $P < 0.05$; **, statistically significant values at a false discovery rate (FDR) of 0.05. ***, statistically significant values under the Bonferroni correction.

[a]Analysis not conducted since there is no orthologous copy in *D. ananassae*.

Table S4. Connections between the insulin/TOR pathway and other pathways in *D. melanogaster*.

IT pathway element	Pathway or molecule	Input/ output	Effect on IT pathway	Kind of evidence	References
p60	Decapentaplegic pathway	input	activation	m	(Higaki and Shimokado 1999; Martin-Castellanos and Edgar 2002)
p110	Ras pathway	input	activation	d	(Orme et al. 2006)
Tsc2 and/or TOR	amino acids	input	activation	d	(Gao et al. 2002)
Rheb	TCTP	input	activation	d	(Hsu et al. 2007)
eIF-4E	LK6	input	activation	d	(Arquier et al. 2005; Reiling et al. 2005)
dMyc	Ras pathway	input	activation	d	(Prober and Edgar 2002)
p60	Susi	input	inhibition	d	(Wittwer et al. 2005)
Tsc1	Scylla/Charybdis	input	inhibition	d	(Reiling and Hafen 2004)
S6K	PP2A	input	inhibition	d	(Bielinski and Mumby 2007)
dMyc	Archipelago	input	inhibition	d	(Moberg et al. 2004)
dMyc	Dco	input	inhibition	d	(Galletti et al. 2007)
dMyc	Wingless pathway	input	inhibition	d	(Johnston et al. 1999; Quinn et al. 2004)
dFOXO	JNK pathway	input	inhibition	d	(Wang et al. 2005)
Shaggy	Hedgehog pathway	input/ output	activation	d	(Jia et al. 2002; Price and Kalderon 2002)
Shaggy	Wingless pathway	input/ output	no effect	d	(Noordermeer et al. 1994)
InR	Dock pathway	output	activation	d	(Song et al. 2003)
PDK1	RSK	output	activation	d	(Rintelen et al. 2001)
S6K	Sima/HIF-1	output	activation	d	(Dekanty et al. 2005)
PKB	Trh	output	activation	d	(Jin et al. 2001)
PKB	sugar metabolism	output	activation	m	(Rulifson et al. 2002)
PKB	lipid metabolism	output	activation	d	(Vereshchagina and Wilson 2006)
PKB	Sima/HIF-1	output	activation	d	(Dekanty et al. 2005)
PKB	ERK pathway	output	activation	d	(Kim et al. 2004)
PKB	apoptosis	output	inhibition	m	(Scanga et al. 2000)

d, interaction observed in *D. melanogaster*; m, interaction observed in mammals that may also exist in *D. melanogaster* as judged from some indirect evidence.

References

Arquier, N., Bourouis, M., Colombani, J., and Leopold, P. 2005. Drosophila Lk6 kinase controls phosphorylation of eukaryotic translation initiation factor 4E and promotes normal growth and development. *Curr Biol* **15:** 19-23.

Avruch, J., Belham, C., Weng, Q., Hara, K., and Yonezawa, K. 2001. The p70 S6 kinase integrates nutrient and growth signals to control translational capacity. *Prog Mol Subcell Biol* **26:** 115-154.

Bai, X., Ma, D., Liu, A., Shen, X., Wang, Q.J., Liu, Y., and Jiang, Y. 2007. Rheb activates mTOR by antagonizing its endogenous inhibitor, FKBP38. *Science* **318:** 977-980.

Bielinski, V.A. and Mumby, M.C. 2007. Functional analysis of the PP2A subfamily of protein phosphatases in regulating Drosophila S6 kinase. *Exp Cell Res* **313:** 3117-3126.

Chou, M.M. and Blenis, J. 1995. The 70 kDa S6 kinase: regulation of a kinase with multiple roles in mitogenic signalling. *Curr Opin Cell Biol* **7:** 806-814.

Dekanty, A., Lavista-Llanos, S., Irisarri, M., Oldham, S., and Wappner, P. 2005. The insulin-PI3K/TOR pathway induces a HIF-dependent transcriptional response in Drosophila by promoting nuclear localization of HIF-alpha/Sima. *J Cell Sci* **118:** 5431-5441.

Dufner, A. and Thomas, G. 1999. Ribosomal S6 kinase signaling and the control of translation. *Exp Cell Res* **253:** 100-109.

Galletti, M., Serras, F., Jiang, J., Pelicci, P.G., Grifoni, D., and Bellosta, P. 2007. In vivo and in vitro regulation of dMyc protein stability by Sgg/dGSK3 and Dco/CK1 kinases. *A. Dros. Res. Conf.* **48:** 121.

Gao, X., Zhang, Y., Arrazola, P., Hino, O., Kobayashi, T., Yeung, R.S., Ru, B., and Pan, D. 2002. Tsc tumour suppressor proteins antagonize amino-acid-TOR signalling. *Nat Cell Biol* **4:** 699-704.

Higaki, M. and Shimokado, K. 1999. Phosphatidylinositol 3-kinase is required for growth factor-induced amino acid uptake by vascular smooth muscle cells. *Arterioscler Thromb Vasc Biol* **19:** 2127-2132.

Hsu, Y.C., Chern, J.J., Cai, Y., Liu, M., and Choi, K.W. 2007. Drosophila TCTP is essential for growth and proliferation through regulation of dRheb GTPase. *Nature* **445:** 785-788.

Jia, J., Amanai, K., Wang, G., Tang, J., Wang, B., and Jiang, J. 2002. Shaggy/GSK3 antagonizes Hedgehog signalling by regulating Cubitus interruptus. *Nature* **416:** 548-552.

Jin, J., Anthopoulos, N., Wetsch, B., Binari, R.C., Isaac, D.D., Andrew, D.J., Woodgett, J.R., and Manoukian, A.S. 2001. Regulation of Drosophila tracheal system development by protein kinase B. *Dev Cell* **1:** 817-827.

Johnston, L.A., Prober, D.A., Edgar, B.A., Eisenman, R.N., and Gallant, P. 1999. Drosophila myc regulates cellular growth during development. *Cell* **98:** 779-790.

Kim, S.E., Cho, J.Y., Kim, K.S., Lee, S.J., Lee, K.H., and Choi, K.Y. 2004. Drosophila PI3 kinase and Akt involved in insulin-stimulated proliferation and ERK pathway activation in Schneider cells. *Cell Signal* **16:** 1309-1317.

Long, X., Lin, Y., Ortiz-Vega, S., Yonezawa, K., and Avruch, J. 2005. Rheb binds and regulates the mTOR kinase. *Curr Biol* **15:** 702-713.

Martin-Castellanos, C. and Edgar, B.A. 2002. A characterization of the effects of Dpp signaling on cell growth and proliferation in the Drosophila wing. *Development* **129:** 1003-1013.

Moberg, K.H., Mukherjee, A., Veraksa, A., Artavanis-Tsakonas, S., and Hariharan, I.K. 2004. The Drosophila F box protein archipelago regulates dMyc protein levels in vivo. *Curr Biol* **14:** 965-974.

Noordermeer, J., Klingensmith, J., Perrimon, N., and Nusse, R. 1994. dishevelled and armadillo act in the wingless signalling pathway in Drosophila. *Nature* **367:** 80-83.

Orme, M.H., Alrubaie, S., Bradley, G.L., Walker, C.D., and Leevers, S.J. 2006. Input from Ras is required for maximal PI(3)K signalling in Drosophila. *Nat Cell Biol* **8:** 1298-1302.

Price, M.A. and Kalderon, D. 2002. Proteolysis of the Hedgehog signaling effector Cubitus interruptus requires phosphorylation by Glycogen Synthase Kinase 3 and Casein Kinase 1. *Cell* **108:** 823-835.

Prober, D.A. and Edgar, B.A. 2002. Interactions between Ras1, dMyc, and dPI3K signaling in the developing Drosophila wing. *Genes Dev* **16:** 2286-2299.

Quinn, L.M., Dickins, R.A., Coombe, M., Hime, G.R., Bowtell, D.D., and Richardson, H. 2004. Drosophila Hfp negatively regulates dmyc and stg to inhibit cell proliferation. *Development* **131:** 1411-1423.

Reiling, J.H., Doepfner, K.T., Hafen, E., and Stocker, H. 2005. Diet-dependent effects of the Drosophila Mnk1/Mnk2 homolog Lk6 on growth via eIF4E. *Curr Biol* **15:** 24-30.

Reiling, J.H. and Hafen, E. 2004. The hypoxia-induced paralogs Scylla and Charybdis inhibit growth by down-regulating S6K activity upstream of TSC in Drosophila. *Genes Dev* **18:** 2879-2892.

Rintelen, F., Stocker, H., Thomas, G., and Hafen, E. 2001. PDK1 regulates growth through Akt and S6K in Drosophila. *Proc Natl Acad Sci U S A* **98:** 15020-15025.

Rulifson, E.J., Kim, S.K., and Nusse, R. 2002. Ablation of insulin-producing neurons in flies: growth and diabetic phenotypes. *Science* **296:** 1118-1120.

Scanga, S.E., Ruel, L., Binari, R.C., Snow, B., Stambolic, V., Bouchard, D., Peters, M., Calvieri, B., Mak, T.W., Woodgett, J.R. et al. 2000. The conserved PI3'K/PTEN/Akt signaling pathway regulates both cell size and survival in Drosophila. *Oncogene* **19:** 3971-3977.

Song, J., Wu, L., Chen, Z., Kohanski, R.A., and Pick, L. 2003. Axons guided by insulin receptor in Drosophila visual system. *Science* **300:** 502-505.

Vereshchagina, N. and Wilson, C. 2006. Cytoplasmic activated protein kinase Akt regulates lipid-droplet accumulation in Drosophila nurse cells. *Development* **133:** 4731-4735.

Wang, M.C., Bohmann, D., and Jasper, H. 2005. JNK extends life span and limits growth by antagonizing cellular and organism-wide responses to insulin signaling. *Cell* **121:** 115-125.

Wittwer, F., Jaquenoud, M., Brogiolo, W., Zarske, M., Wustemann, P., Fernandez, R., Stocker, H., Wymann, M.P., and Hafen, E. 2005. Susi, a negative regulator of Drosophila PI3-kinase. *Dev Cell* **8:** 817-827.

3.1.4. Addenda

3.1.4.1. Distribució dels gens de la via de la insulina/TOR al genoma de *D. melanogaster*.

Per tal de contrastar si els gens implicats en la via de transducció de senyal de la insulina/TOR es distribueixen aleatòriament al genoma de *D. melanogaster*, es va emprar el mètode de Monte Carlo descrit per Lee i Sonnhammer (2003). El mètode es basa en el càlcul del *clustering score* (*cs*), que és la suma dels *pair scores* (*ps*) per a tots els parells de gens possibles:

$$cs = \sum_{i>j} ps_{ij} \ , \qquad (1)$$

on *i* i *j* són els subíndexs corresponents als dos gens. Els ps_{ij} es calculen de manera diferent segons si els dos gens es troben o no en el mateix cromosoma, essent:

$$ps_{ij} = \begin{cases} d/g, & \text{per a gens ubicats en el mateix braç cromossòmic,} \\ d/m, & \text{en cas contrari;} \end{cases} \qquad (2)$$

on *d* és la longitud mitjana de tots els braços cromosòmics del genoma, *g* és la distància entre els dos gens al cromosoma, i *m* és la mitjana de les longituds dels cromosomes en els quals es troben els gens.

La significació estadística del *clustering score* es va determinar en base a 10.000 pseudorèpliques del mateix nombre de gens que en el conjunt original obtinguts aleatòriament del genoma de *D. melanogaster* (*release* 5.1). Atès que fer servir grups de gens paràlegs podrien esbiaixar les anàlisis, aquests es van realitzar de manera separada per cada un dels paràlegs de *eIF-4E*. Així, en cada cas es va analitzar la distribució d'un total de 21 gens.

Els resultats no permeten descartar una distribució aleatòria dels gens de la via de la insulina/TOR ($P \geq 0{,}0686$; independentment del paràleg de *eIF-4E* triat). Això sembla anar en contra de que s'havia proposat que els gens implicats en aquesta via no presentarien una distribució aleatòria al genoma de *D. melanogaster* (De Jong i Bochdanovits 2003). Una anàlisi a escala genòmica de la distribució dels gens implicats en diferents vies (tant de transducció de senyal com metabòliques) va mostrar que aquests no es distribueixen a l'atzar,

sinó que els gens d'una mateixa via tenen certa tendència a pressentar-se agrupats als genomes de diferents espècies, incloent-hi *D. melanogaster* (Lee i Sonnhammer 2003). Cal remarcar, però, que dels cinc genomes analitzats, el de *D. melanogaster* és el que presenta un menor grau d'agrupació, cosa que s'ha atribuït a l'elevada taxa d'evolució cromosòmica dels genomes de *Drosophila* (superior a la d'altres eucariotes; Ranz et al. 2001).

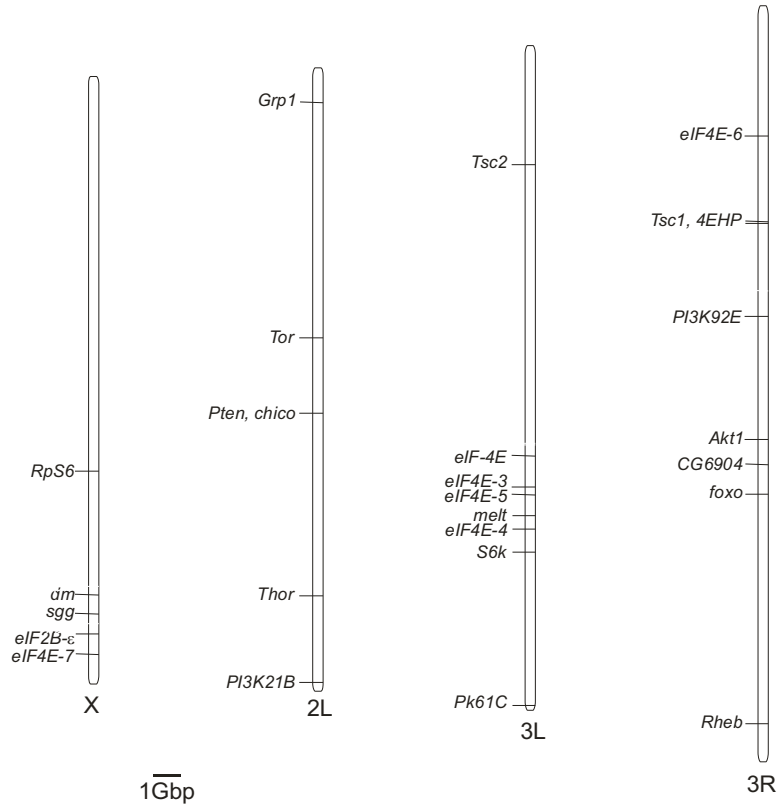

Figura A1.
Distribució cromosòmica dels gens implicats en la via de transducció de senyal de la insulina/TOR al genoma de *D. melanogaster*.

3.1.4.2. Nivells de limitació funcional a variar als paràlegs de *eIF-4E*

Tot i que 7 dels gens estudiats són paràlegs entre si, difereixen considerablement en els seus nivells de limitació funcional a variar. Així, analitzant els patrons de divergència de les 5 espècies del subgrup *melanogaster* (ja que *eIF4E-6* només compta amb ortòlegs en aquest grup d'espècies), els valors de ω oscil·len entre 0,073 (per a *eIF4E-4*) i 0,576 (*eIF4E-6*). Existeix un estudi funcional d'aquests gens a *D. melanogaster* (Hernandez et al. 2005), segons el qual els productes de 5 d'aquestes gens (tots excepte *eIF4E-6* i *4EHP*) són capaços d'interactuar amb eIF4G i de rescatar el creixement cel·lular de llevats mutants deficients en eIF4E. En aquest treball es va concloure que els productes de *eIF4E-6* i *4EHP*, que són els únics amb diferències aminoacídiques no conservatives en els residus rellevants per a la interacció amb eIF4E, podrien ser reguladors negatius de la traducció o bé proteïnes no funcionals. Existeix, doncs, una certa correspondència entre les nostres estimes de ω amb les conclusions d'aquest darrer estudi. De fet, el valor de ω per a *eIF4E-6* és, amb diferència, el més elevat dels observats en aquests gens. A més, en l'estudi funcional es mostra com una de les isoformes codificades per *eIF-4E* (la triada en el present estudi) és la que s'expressa de manera majoritària durant el cicle biològic de *D. melanogaster*, la qual cosa és consistent amb el fet que el valor de ω per a aquest gen és el segon més baix (i del mateix ordre que el valor més baix).

Taula A1. Valors de ω per al grup de paràlegs de *eIF-4E*

Gen	ω
eIF-4E	0,077
eIF4E-3	0,254
eIF4E-4	0,073
eIF4E-5	0,256
eIF4E-6	0,576
eIF4E-7	0,345
4EHP	0,155

3.1.4.3. Informació complementària sobre les entrades i sortides de la via de la insulina/TOR a *D. melanogaster*

Entrades

p60

La proteïna Susi regula negativament l'activitat de PI3K tot unint-se directament a la seva subunitat p60. El gen *Susi* s'expressa d'acord amb un cicle circadià, amb nivells d'expressió més elevats a la nit (Claridge-Chang et al. 2001; McDonald i Rosbash 2001; Ueda et al. 2002).

A mamífers, les proteïnes PI3K, PDK1 i PKB són regulades positivament per la via TGF-β1 (Higaki i Shimokado 1999). A l'ala de *D. melanogaster*, Decapentaplegic (Dpp, l'homòleg de TGF-β) afecta el creixement i la proliferació de manera dependent de PI3K (Martin-Castellanos i Edgar 2002). Tot i que aquestes observacions són compatibles amb la possibilitat de que dPI3K sigui activada per la via Dpp, aquesta interacció encara no ha estat demostrada (Goberdhan i Wilson 2003).

p110

L'activació de Dp110 mediada per Ras és indispensable en algunes situacions per a assolir la màxima activació de Dp110, com ara la producció d'ous.

Tsc1

Scylla i Charybdis regulen positivament el complexe TSC, inhibint per tant la via de la insulina/TOR. La transcripció dels gens *Scyl* i *Char* és induïda en condicions d'hipòxia, presumiblement per part de Sima.

Complex TSC i/o TOR

L'activitat de TOR depèn de la disponibilitat d'aminoàcids, tot i que no està clar si l'element sensible als aminoàcids és el complex TSC, TOR o tots dos.

Rheb

La proteïna dTCTP s'associa directament amb dRheb i actua com un intercanviador de guanina d'aquesta proteïna.

S6K

S6K és fosforilada al lloc de fosforilació de TOR per la fosfatasa PP2A.

dFOXO

La via JNK s'activa en determinades situacions d'estrès com ara la radiació UV o l'estrès oxidatiu. L'activació d'aquesta via promou la localització nuclear de dFOXO.

eIF4E

eIF4E s'activa quan és fosforilat per la cinasa Lk6, que, al seu torn, s'activa en resposta a mitògens, citoquines i condicions d'estrès cel·lular

Shaggy

La via Wingless inactiva Shaggy, la qual cosa promou l'activació de Armadillo. Almenys a mamífers, la transmissió dels senyals d'ambdues vies a través de GSK-3 (proteïna homòloga a Shaggy) semblen ésser independents: la via Wingless, per tant, sembla estar aïllada de la via de la insulina (Ding et al. 2000).

Shaggy sembla estar implicada en la via Hedgehog, i fosforila el factor de transcripció Cubitus interruptus, tot promovent-ne la proteòlisi.

dMyc

Archipelago (Ago) s'uneix a dMyc, tot promovent-ne la degradació. La transcripció de Fbw7/hCDC4 (l'orgòleg humà de *Ago*) és activada per p53 (Kimura et al. 2003), que al seu torn és activada per diferents estressos cel·lulars, tot incloent un augment de temperatura, condicions d'hipòxia, shock osmòtic i danys al DNA.

La via Ras inhibeix la degradació de dMyc mitjançant la via Raf/MAPK, de manera independent de la via PI3K.

dMyc és fosforilat per Dco, la proteïna ortòloga a *D. melanogaster* de la cinasa de caseïna 1, cosa que promou la ubiquitinització i degradació de dMyc.

La via Wingless inhibeix l'expressió del gen *dmyc* i la funció del seu producte proteïc.

Sortides

InR

InR activa la via Dock, la qual cosa permet als axons de les cèl·lules fotoreceptores trobar el camí des de la retina fins el cervell durant el desenvolupament. La interacció InR—Dock té lloc mitjançant l'extensió C-terminal de InR (present a *Drosophila* i no a mamífers).

PDK1

L'activitat de RSK és modulada per PDK1, probablement mitjançant fosforilació directa.

S6K i PKB

S6K i PKB activen la proteïna Sima, la qual cosa indueix la transcripció de *Scyl*. Scylla, al seu torn, activa el complex TSC, la qual cosa resulta en un bucle de retroacció negativa.

PKB

La fosforilació de Trh per PKB a la Serina 665 és essencial per a la localització nuclear i l'activació funcional d'aquesta proteïna.

PKB està implicada en l'activació de la via ERK induïda per la insulina, la qual cosa porta a la proliferació cel·lular a *D. melanogaster*.

Als mamífers, PKB inactiva el factor pro-apoptòtic Bad (Datta et al. 1997). A *D. melanogaster*, PKB és essencial per a la supervivència cel·lular (Scanga et al. 2000), tot i que les vies que medien aquests efectes encara no estan establertes (Goberdhan i Wilson 2003).

Als mamífers, PKB regula la incorporació de glucosa a la cèl·lula. Tot i que a *D. melanogaster* la via de la insulina té un paper clau en el metabolisme dels sucres, les bases moleculars d'aquesta regulació són encara desconegudes.

PKB està també implicada en el metabolisme lipídic en les cèl·lules nodrissa de l'ovari a *D. melanogaster*.

Referències Addenda Article 1

Claridge-Chang A et al. 2001. Circadian regulation of gene expression systems in the Drosophila head. Neuron. 32:657-671.

Datta SR et al. 1997. Akt phosphorylation of BAD couples survival signals to the cell-intrinsic death machinery. Cell. 91:231-241.

De Jong G, Bochdanovits Z. 2003. Latitudinal clines in Drosophila melanogaster: body size, allozyme frequencies, inversion frequencies, and the insulin-signalling pathway. J Genet. 82:207-223.

Ding VW, Chen RH, McCormick F. 2000. Differential regulation of glycogen synthase kinase 3beta by insulin and Wnt signaling. J Biol Chem. 275:32475-32481.

Goberdhan DC, Wilson C. 2003. The functions of insulin signaling: size isn't everything, even in Drosophila. Differentiation. 71:375-397.

Hernandez G et al. 2005. Functional analysis of seven genes encoding eight translation initiation factor 4E (eIF4E) isoforms in Drosophila. Mech Dev. 122:529-543.

Higaki M, Shimokado K. 1999. Phosphatidylinositol 3-kinase is required for growth factor-induced amino acid uptake by vascular smooth muscle cells. Arterioscler Thromb Vasc Biol. 19:2127-2132.

Kimura T, Gotoh M, Nakamura Y, Arakawa H. 2003. hCDC4b, a regulator of cyclin E, as a direct transcriptional target of p53. Cancer Sci. 94:431-436.

Lee JM, Sonnhammer EL. 2003. Genomic gene clustering analysis of pathways in eukaryotes. Genome Res. 13:875-882.

Martin-Castellanos C, Edgar BA. 2002. A characterization of the effects of Dpp signaling on cell growth and proliferation in the Drosophila wing. Development. 129:1003-1013.

McDonald MJ, Rosbash M. 2001. Microarray analysis and organization of circadian gene expression in Drosophila. Cell. 107:567-578.

Ranz JM, Casals F, Ruiz A. 2001. How malleable is the eukaryotic genome? Extreme rate of chromosomal rearrangement in the genus Drosophila. Genome Res. 11:230-239.

Scanga SE et al. 2000. The conserved PI3'K/PTEN/Akt signaling pathway regulates both cell size and survival in Drosophila. Oncogene. 19:3971-3977.

Ueda HR et al. 2002. Genome-wide transcriptional orchestration of circadian rhythms in Drosophila. J Biol Chem. 277:14048-14052.

3.2. Article 2
Comparative genomics of the vertebrate insulin/TOR signal transduction pathway genes: A network-level analysis of selective pressures along the pathway

David Alvarez-Ponce, Montserrat Aguadé i Julio Rozas

(en preparació)

3.2.1. Resum

La complexitat de la funció biològica es basa en grans xarxes de molècules que interactuen entre si. Les propietats evolutives d'aquestes xarxes, però, disten considerablement d'ésser enteses completament. S'ha mostrat que les pressions selectives depenen de la posició que els gens ocupen dins la xarxa. Anteriorment, vam mostrar que a la via de transducció de senyal de la insulina/TOR de *Drosophila* hi ha una correlació entre la posició a la via i la intensitat de la selecció purificadora, essent els gens que actuen a la part final els més limitats a variar. Aquí hem estudiat la dinàmica evolutiva d'aquesta via ben caracteritzada als vertebrats. Més concretament, hem estudiat l'impacte de la selecció natural sobre l'evolució de 72 gens d'aquesta via. Hem trobat que als vertebrats també hi ha un gradient similar en els nivells de limitació funcional al llarg de la via de la insulina/TOR. Aquest patró no és el resultat d'una polaritat en l'impacte de la selecció positiva ni d'una sèrie de factors que afecten el grau limitació funcional (nivell i rang d'expressió gènica, biaix en l'ús de codons, longitud de les proteïnes i connectivitat. També vam observar que els gens de la via que codifiquen proteïnes que interactuen físicament evolucionen sota nivells de limitació funcional semblants. Els resultats indiquen que l'arquitectura de la via de la insulina/TOR dels vertebrats constreny l'evolució molecular dels seus components. Per tant, la polaritat detectada a *Drosophila* no és específica ni incidental d'aquest gènere. Així doncs, tot i que els mecanismes biològics subjacents romanen sense establir, aquests probablement siguin similars als vertebrats i a *Drosophila*.

Article 2

Comparative genomics of the vertebrate insulin/TOR signal transduction pathway genes: A network-level analysis of selective pressures along the pathway

David Alvarez-Ponce, Montserrat Aguadé and Julio Rozas

2010

(En preparació)

Departament de Genètica, Facultat de Biologia, Universitat de Barcelona, Barcelona 08028, Spain.

ABSTRACT

The complexity of the biological function relies on large networks of interacting molecules. The evolutionary properties of these networks, however, are far from fully understood. It has been shown that selective pressures depend on the position of genes in the network. We have previously shown that in the *Drosophila* insulin/TOR signal transduction pathway there is a correlation between pathway position and the strength of purifying selection, being downstream genes the most constrained. Here we have studied the evolutionary dynamics of this well-characterized pathway in vertebrates. More specifically, we have studied the impact of natural selection on the evolution of 72 genes of this pathway. We have found that in vertebrates there is also a similar gradient on the selective constraint levels along the insulin/TOR pathway. This feature is neither the result of a polarity in the impact of positive selection nor of a series of factors affecting selective constraint (gene expression level and breadth, codon bias, protein length, and connectivity). We also found that pathway genes encoding physically interacting proteins tend to evolve under similar selective constraints. The results indicate that the architecture of the vertebrate insulin/TOR pathway constrains the molecular evolution of its components. Therefore, the polarity detected in *Drosophila* is neither specific nor incidental of this genus. Hence, although the underlying biological mechanisms remain unclear, they may be similar in both vertebrates and *Drosophila*.

INTRODUCTION

The neutral theory of molecular evolution predicts an inverse correlation between the functional significance of genome regions and polymorphism and divergence (Kimura 1983). Indeed, the levels and patterns of purifying selection widely vary across different genes and genomic regions. The evolutionary meaning of such variation is a major topic in evolutionary biology. There are a number of factors affecting selective constraint levels acting on genes, including expression level and breadth (Duret and Mouchiroud 2000; Pál et al. 2001; Subramanian and Kumar 2004), codon bias (Sharp 1991; Pál et al. 2001), protein length (Subramanian and Kumar 2004), or molecular function (e.g., Castillo-Davis et al. 2004). These factors, however, account for only a reduced fraction of the variation in selective constraint levels, and particularly in higher eukaryotes (e.g., Ingvarsson 2007).

The role of natural selection in the evolution of biological complex systems is poorly understood (Cork and Purugganan 2004). Yet, genes do not act in isolation, but interact with many other components in complex networks. The recent availability of large-scale protein–protein interaction and metabolic information allows studying the impact of a gene's position in a network on its pattern of evolutionary change. Remarkably, elements with a higher connectivity or centrality in the network tend to be highly constrained (Fraser et al. 2002; Hahn and Kern 2005), while physically interacting proteins show correlated evolutionary histories (e.g., Fryxell 1996; Fraser et al. 2002). These observations clearly indicate that network architecture constrains the molecular evolution of its components.

There is also compelling evidence linking network position and evolutionary change in specific and well characterized pathways. Specific enzymes in a pathway can contribute differentially to the overall pathway function (and, hence, to the associated phenotypes). Genes encoding enzymes with high control coefficients (those exerting a

relatively high influence over flux; Kacser and Burns 1973), such as those acting at network branch points (LaPorte et al. 1984; Stephanopoulos and Vallino 1991), are expected to evolve under strong natural selection (Hartl et al. 1985; Eanes 1999; Watt and Dean 2000). Consistently, in the glucose catabolic pathways of *Drosophila* positive selection has acted preferentially on genes encoding branch point enzymes (Flowers et al. 2007). Furthermore, it has been proposed that genes acting in the upstream part of pathways evolve under stronger purifying selection than those in the downstream part, since mutations at these genes may have more pleiotropic effects. Consistently, Rausher and co-workers found that, in the plant anthocyanin biosynthetic pathway, selective constraint levels correlate with the position along the upstream/downstream axis of the pathway, being upstream genes the most constrained (Rausher et al. 1999). This polarity seems to be neither explained by differences in mutation rates (Lu and Rausher 2003), nor by the action of positive selection (Rausher et al. 2008) along the pathway. A similar polarity in the selective constraint distribution along the pathway has been observed in the plant isoprene, terpenoid, and carotenoid biosynthetic pathways (Sharkey et al. 2005; Livingstone and Anderson 2009; Ramsay et al. 2009), and in the *Drosophila* Ras signaling pathway (Riley et al. 2003). This feature, nevertheless, is not general (Olsen et al. 2002; Jovelin et al. 2009; Yang et al. 2009), but rather may depend on the particular pathway architecture. Indeed, we showed that in the insulin/TOR (IT) signal transduction pathway of *Drosophila* this polarity occurs in the opposite direction, i.e., purifying selection is higher in the downstream genes (Alvarez-Ponce et al. 2009).

The IT signal transduction pathway plays a central role in multiple fundamental biological processes, as diverse as growth, energetic metabolism, reproduction and aging (Oldham and Hafen 2003; LeRoith et al. 2004; Taguchi and White 2008). Additionally, diseases such as insulin resistance, diabetes, obesity and cancer are associated with

dysregulation of some genes of the pathway. This pathway has been well characterized in a series of organisms, and both its structure and function is highly conserved from insects to vertebrates. This molecular pathway, therefore, offers a good opportunity to study the relationship between the pathway architecture and gene sequence evolution across a wide range of phylogenetic groups.

Here, we study the vertebrate IT signal transduction pathway to establish whether the polarity in the strength of purifying selection observed in *Drosophila* is incidental or something specific to this genus or, on the contrary, it represents a more general feature. For this purpose, we have characterized the molecular evolution of the IT pathway genes in the complete genomes of 6 vertebrates. We have identified, and manually annotated, the orthologs and paralogs of 72 genes involved in the human IT pathway and have reconstructed their evolutionary history. We found that, as previously observed in *Drosophila*, downstream genes are the most constrained in the vertebrate IT pathway. Therefore, the polarity distribution of selective constraints along the pathway is neither incidental nor specific of the *Drosophila* genus, and would reflect the action of a more general biological mechanism.

METHODS

Selection of IT pathway genes for the analysis

We defined the set of genes encoding the human IT signal transduction pathway to be analyzed by searching the literature for known human orthologs of those genes included in our previous analysis of the *Drosophila* IT pathway (Alvarez-Ponce et al. 2009). We also included in this set the insulin receptor gene (*INR*), its closest paralogs encoding the IGF1 receptor (*IGF1R*) and the insulin receptor-related receptor (*INSRR*), and the 9 Protein Kinase C (PKC)-encoding genes. We also studied the closest annotated paralogs of the selected genes (Ensembl database version 50; Flicek et al. 2008).

We tried to identify unannotated paralogs by using a two-round BLAST search. First, for each human protein we performed a TBLASTN search (E-value < 10^{-5}) against the human genome (International Human Genome Sequencing Consortium 2004). Second, the TBLASTN hits were used as query in a BLASTP search against the human proteome. If the best hit corresponded to the original gene or one of its paralogs, with a sequence identity higher than 60%, and covering at least 50% of the sequence length, we manually annotated this sequence and included it in the analyses. The final set of genes (table S1) consists of 72 genes (belonging to 24 paralogous groups) and 43 pseudogenes (40 are intronless, likely processed copies).

Identification and annotation of IT pathway genes in nonhuman vertebrates

We searched the IT pathway genes in the genome sequences of the mammals *Mus musculus* (Mouse Genome Sequencing Consortium 2002), *Bos taurus* (The Bovine Genome Sequencing and Analysis Consortium 2009), *Monodelphis domestica* (Mikkelsen et al. 2007) and *Ornithorhynchus anatinus* (Warren et al. 2008), and the bird *Gallus gallus* (International Chicken Genome Sequencing Consortium 2004). We retrieved the coding

sequences (CDSs) of the human genes and their predicted orthologs from the Ensembl database. For genes with alternative splicing, we chose the variant encoding the longest protein among those shared across the six species (table S1).

Given that the Ensembl information is mainly based on computational gene predictions, we visually inspected and, when required, manually reannotated all sequences. For that purpose, we (1) removed exons without correspondence with the human orthologs; (2) added exons missing in the original dataset; and (3) merged separated gene model predictions that were different portions of the same gene. We also searched the GenBank database for incomplete or missing genes.

We performed a two-round BLAST search to identify nonhuman unannotated sequences. Each human protein was used as query in a TBLASTN search (E-value < 10^{-5}) against all nonhuman genomes, and the resulting hits were used as query in a second TBLASTN search against the human genome. Sequences giving as best hit either the original gene or one of its paralogs were manually annotated and included in the analyses.

Sequences with premature stop codons or frameshifts were classified as putative pseudogenes. We confirmed these features by inspecting the corresponding trace archives. If an individual sequencing read did not contain the disrupting feature, or if all concerned chromatograms had low quality at the affected positions, these features were considered as sequencing errors. We also examined the trace archives to determine whether some paralogous copies were in fact the result of erroneous genome assembly due to sequencing errors. For that purpose, we checked the quality of the affected sequencing traces at the mismatch positions, and all groups of putative paralogs with no confirmed differences were considered as a single copy.

Multiple sequence alignment and phylogenetic analysis

We inferred the orthology/paralogy relationships by phylogenetic analysis. For that purpose, we generated a multiple sequence alignment (MSA) of proteins for each homology group using Probcons 1.11 (Do et al. 2005). These alignments were used to guide the alignment of the CDS sequences. We built a neighbor-joining tree for each MSA based on either the CDS or the protein sequences (in function of the divergence level), using the software MEGA4 (Tamura et al. 2007) and applying either the Tamura-Nei (Tamura and Nei 1993) or the JTT (Jones et al. 1992) evolution models.

We generated a separate MSA for each orthologous group. We excluded from these alignments all sequences with any pseudogenic feature, and only groups with putatively functional representatives in all six species were considered. For those orthologous groups with multiple copies in a given genome (i.e., co-orthologs), we used the sequence that covers the largest fraction of the human ortholog. If two orthologous groups shared a particular sequence (due to gene duplication after the mammals/birds split), we chose only the orthologous group most directly involved in the IT pathway according to the literature to avoid redundancy. All MSAs were manually curated using the software BioEdit 7.0.5.2 (Hall 1999), and poorly alignable positions were discarded from the analyses.

We evaluated the impact of natural selection on gene evolution from the nonsynonymous (d_N) to synonymous (d_S) divergence ratio ($\omega = d_N/d_S$). Values of ω lower than 1 indicate the action of purifying selection, whereas $\omega = 1$ and $\omega > 1$ are indicative of strictly neutral and adaptive evolution, respectively. We obtained ω estimates by applying two evolutionary models implemented in the codeml program from the PAML 3.15 package (Yang 1997). The M0 model assumes a single ω value across all codons and phylogenetic branches, whereas the free-ratio (FR) model assumes an independent ω value

for each branch. We tested for the presence of codons evolving under positive selection, contrasting the M1a and M2a models (Wong et al. 2004), and the M7 and M8 models (Yang et al. 2000) by the likelihood ratio test (Whelan and Goldman 1999). A significantly better fit to the data of models M2a or M8 was interpreted as evidence of positive selection. We controlled for the false discovery rate (FDR) associated with multiple testing at $q = 0.05$ (Benjamini and Hochberg 1995). We used the Bayes Empirical Bayes approach (Yang et al. 2005) to identify the specific codons evolving under positive selection (posterior probability ≥ 95%). All codon-based analyses were conducted using the accepted species tree topology (figure 1), the F3×4 codon frequency model (Goldman and Yang 1994), and three different starting ω values (0.01, 0.1 and 1) to overcome the multiple local optima problem. Any set of FR estimates (d_N, d_S and ω) with $\omega > 3$, $d_S > 5$, or $S \times d_S < 1$ (where S is the number of synonymous positions) were discarded from the analyses.

Network-level analysis

The structure of the IT pathway (biochemical information extracted from the literature) was encoded into a directed graph (termed *G*; figure 2A) with nodes and arcs representing proteins and activatory/inhibitory interactions, respectively. This graph consists of 21 nodes connected by 39 arcs, of which 32 are physical (direct protein–protein interactions, PPIs; figure 2B). We used this graph to assign the position of each pathway element, computed as the number of steps required to transduce the signal from the insulin/IGF1 receptor (position 0) to the remaining elements in the pathway (the maximum was 10 steps). Paralogous genes share the same pathway position. Nevertheless, paralogous copies not involved in insulin signaling (*INSRR*, *PIK3CG*, *EIF4E2* and *EIF4E3*) were eliminated from the network-level analyses. In the end, a total of 58 genes

have assigned pathway position, but only 48 of them have copies in all six species and were therefore used in network-level analyses.

We contrasted whether physically interacting IT pathway proteins tend to exhibit similar d_N, d_S or ω values by applying a Monte Carlo method (Fraser et al. 2002). For this analysis, we used a subgraph of *G* containing only physical interactions (denoted as *S*; figure 2B), and the average absolute difference between the d_N, d_S or ω values of pairs of physically interacting elements in the IT pathway (*X*) as statistic:

$$X = \frac{1}{n}\sum_{i=1}^{n}\left|x_{i1} - x_{i2}\right|,$$

where $n = 32$ is the number of interacting pairs in *S*, and x_{i1} and x_{i2} are the d_N, d_S or ω values of the two genes encoding interacting proteins (1 and 2) at pair *i*. We contrasted whether *X* is significantly lower than that expected at random by generating 100,000 randomizations of *S*. Each random network has the same 21 nodes and the same number of arcs ($n = 32$) connected by two different nodes (sampled without replacement). *P*-values were computed as the proportion of simulated networks having *X* values lower than the observed. We conducted a modification of this Monte Carlo test controlling for the association between pathway position and selective constraint. For that purpose, we applied linear regression to model the relationship between pathway position and either ω or d_N, and used the residuals of the model (i.e., the difference between observed and predicted values) for the Monte Carlo analysis.

Statistical tests of association

We used the non-parametric Spearman's rank correlation coefficient (ρ) to contrast whether d_N, d_S and ω estimates correlate with pathway position along the IT pathway. Because these parameters are affected by a number of factors, including gene expression level and breadth (Duret and Mouchiroud 2000; Pál et al. 2001; Subramanian and Kumar

2004), codon bias (Sharp 1991; Pál et al. 2001), protein length (Subramanian and Kumar 2004), and connectivity (Fraser et al. 2002), a polarity in these factors along the upstream/downstream IT pathway axis could potentially account for the distribution of d_N, d_S and ω. Therefore, we included all these factors in order to factor out their effect on sequence evolution. The information was gathered from multiple sources:

- *Expression level and breadth*: We used the human expression data from Su et al. (2004) (U133A+GNF1H dataset, normalized using the MAS5 algorithm), which contains gene expression measures for 79 different tissues (or organs) with two replicates each. We excluded data from cancerous tissues given that IT pathway elements are usually dysregulated in cancer conditions. Furthermore, since some organs are represented by multiple entries in the dataset (for instance, the brain is represented by multiple entries, including the whole brain and different parts of it), we only used a set of 25 nonredundant tissues (table S2) to avoid biasing the results. For each gene and tissue, values were averaged across the two replicates. When multiple probes matched the same gene, we chose that with the highest average signal. For each gene, the expression level was estimated as the average across the 25 selected tissues, while expression breadth as the number of tissues where it is expressed (expression level ≥ 200; see Su et al. 2002).
- *Connectivity*: We obtained protein–protein interaction data from the human interaction network of Bossi and Lehner (2009). This dataset consists of 80,922 physical interactions gleaned from 21 different sources, of which 2030 involve IT pathway components. The connectivity of each gene was computed as the number of PPIs in which it is involved.
- *Codon bias*: For each orthologous group, the codon bias was estimated as the median of the effective number of codons (ENC; Wright 1990) across all six studied species.

ENC values were computed using the DnaSP 5.00.02 software (Librado and Rozas 2009).

- *Protein length*: Since many nonhuman sequences are incomplete, and protein length is highly conserved across species (Wang et al. 2005), we used the human value.

We conducted a bivariate correlation analysis among these factors, the pathway position, and the d_N, d_S and ω estimates. Furthermore, we applied two multivariate analysis techniques (path analysis and partial correlation) to better characterize the relationships among these factors. For path analysis, we used the causal model depicted in figure 3 and, when needed, variables were either log- or square root-transformed to improve normality. We conducted these analyses using the packages AMOS 17 (for path analysis), PASW Statistics 17 (for bivariate correlation analysis), and R (Ihaka and Gentleman 1996) (for partial correlation analysis). Throughout this paper, we report two-tailed *P*-values except for the association between pathway position and d_N and ω, where we have an *a priori* hypothesis about the direction of the association (one-tailed tests).

We used three different datasets for most analyses:

- *Dataset 1*: This dataset includes the 48 genes used in network-level analysis (elements with assigned pathway position and present in all six species; table S3).
- *Dataset 2*: This dataset is a subset of dataset 1 that includes a single gene per paralogous group (n = 21; table 1). We used this dataset to avoid using information on multiple paralogous copies, which may exhibit similar selective constraint levels and are, therefore, not suitable for correlation analyses (all observations should be independent). We chose a single paralog per group according to available information on the paralogs molecular function (obtained from the literature). We chose, among the copies present in all six studied species, (1) the copy that plays a more direct role in the IT pathway (e.g., the copy where mutations affect more severely insulin signaling);

(2) the copy whose activation is more affected by insulin signaling; (3) the embryonic lethal paralog; or (4) the archetypical copy performing all functions (only partially covered by its paralogs). In those cases with incomplete information on the differential function of paralogs, we chose the copy with a higher expression breadth.

- *Dataset 3*: In this dataset, also derived from dataset 1, values were averaged across all copies for each paralogous group (n = 21; table S4).

RESULTS

Distribution of insulin/TOR pathway genes across vertebrates

We applied a combination of automatic and manual methods to identify and annotate the orthologs of 115 IT pathway human sequences (72 genes and 43 pseudogenes; table S1) in the other vertebrate genomes. We identified 617 putative orthologs of the human genes (332 putatively functional genes, 246 pseudogenes and 39 intronless sequences; table S5). The current analysis, therefore, encompasses a total of 734 sequences. Since current genome data comprise unsequenced regions, this number should be considered as the minimum number of sequences (genes plus pseudogenes). Moreover, recent duplicates might have been treated as a single copy during genome assembly. It should be noted, however, that the six genomes have high-coverage sequence data (from 6X to 10X) and, therefore, all putatively missing genes likely are truly absent. Interestingly, we did not identify any pseudogene nor processed copy in the chicken genome, which agrees with the small number of processed copies detected in this genome [51 (International Chicken Genome Sequencing Consortium 2004), in contrast with the more than 15,000 detected in mammalian genomes (Torrents et al. 2003; Rat Genome Sequencing Project Consortium 2004)].

Two hundred and thirty-eight (out of 734) sequences belong to the ribosomal protein S6 (*RPS6*) homology group (6 genes, 212 pseudogenes and 20 intronless sequences; table S5). This is in agreement with previous observations in mammalian genomes showing that each ribosomal protein (RP) is encoded by only a single gene with introns that has several processed pseudogenes. Indeed, over 2400 RP processed pseudogenes have been identified in the human genome, in contrast with 79 functional copies (Zhang et al. 2002). Consistent with our observations, multiple processed *RPS6*

pseudogenes have been described in both the human and mouse genomes (Antoine and Fried 1992; Feo et al. 1992; Pata and Metspalu 1996; Zhang et al. 2002).

Sixty (out of 72) genes have putative functional copies in all genomes, and all paralogous groups have at least one nonpseudogenic copy in each genome. Therefore, the function of missing genes may be undertaken by some of their functional paralogs. Our results, therefore, suggest that all species have a complete IT pathway.

Impact of natural selection on gene sequence evolution

Estimates of ω (under the M0 model) range from 0.002 (for *GSK3B* and *RPS6* genes) to 0.140 (for *TSC1*), with a median value of 0.116 (table S3). These values indicate that the IT pathway genes are under relatively strong purifying selection, suggesting that all genes are functional. We performed two maximum likelihood tests for positive selection (table S6). Although there were no significant results in the M2a vs. M1a comparison, the M8 vs. M7 test allowed identifying 3 genes exhibiting the molecular signature of positive selection: *IRS4*, *AKT3* and *PRKCD* ($P < 0.05$). None of these results, however, is significant after controlling for the FDR. Therefore, positive selection would not be a major force driving the evolution of the IT pathway genes in vertebrates.

Relationship between the selective constraint of interacting proteins

We used a Monte Carlo approach to contrast whether genes encoding physically interacting proteins (figure 2B) tend to evolve under similar selective constraint levels (Fraser et al. 2002). Because current knowledge on the interactions among proteins encoded by different paralogous copies is very incomplete, we restricted the analysis to datasets 2 (containing a single gene per paralogous group; table 1) and 3 (where values are averaged across paralogs; table S4). We found that ω values of genes encoding physically

interacting proteins are more similar than expected from a random network (dataset 2: X_ω = 0.023, P = 0.0025; dataset 3: X_ω = 0.024, P = 0.0122; table S7). The analysis conducted separately for d_N and d_S yields significant results for d_N (dataset 2: X_N = 0.079, P = 0.0050; dataset 3: X_N = 0.084, P = 0.0028; table S7), but not for d_S (dataset 2: X_S = 1.983, P = 0.873; dataset 3: X_S = 1.316, P = 0.702; table S7). These results would indicate that amino acid changes are the main contributors to the observed similarity in selective constraint values between interacting proteins.

Levels of selective constraint along the IT pathway

We tested whether there exists a polarity in the selective constraint levels (estimated from the ω and d_N values) along the upstream/downstream IT pathway axis. Although nonsignificant, the sign of the correlation between pathway position and ω is negative for all three datasets (dataset 1: Spearman's rank correlation coefficient, ρ = –0.073, P = 0.312; dataset 2: ρ = –0.136, P = 0.279; dataset 3: ρ = –0.134, P = 0.281; tables 2, S8 and S9, figure 4). A similar trend is observed for the d_N values; that is, the tests are not significant, but the sign of the correlation is always negative (tables 2, S8 and S9).

We also conducted the correlation analysis separately in each phylogenetic branch (figure 5; tables 3, S10 and S11). For datasets 2 and 3, the correlation between ω, or d_N, and pathway position is negative in all 9 branches (a number significantly higher than the 50% expected at random; binomial test, P = 0.002). For dataset 1, the correlation between ω and the pathway position is negative in 7 branches, which does not represent a significant departure from 50% (binomial test, P = 0.090), whereas the correlation between d_N and pathway position if negative in 8 branches (binomial test, P = 0.020). Furthermore, the correlation between pathway position and ω is significant in two branches (regardless of the dataset used), whereas for d_N the correlation is significant for either two (datasets 1

and 3) or four branches (dataset 2). The direction of the correlation between pathway position and d_S is negative either in 7 (datasets 1 and 2) or 6 branches (dataset 3), which does not represent a significant departure from 50% (binomial test, $P = 0.090$, $P = 0.254$, respectively). This correlation is significantly negative for either one (dataset 1) or two branches (datasets 2 and 3).

Because the available genome sequence data of *O. anatinus* is highly fragmented, we re-evaluated the former correlations removing this species from the analyses. This involved an average increase of 11.13% in the number of analyzed codons. Remarkably, this reanalysis uncovers a significant correlation between d_N and pathway position for dataset 2 ($\rho = -0.441$, $P = 0.023$; table 2). The correlation between pathway position and ω is negative in either all 7 (datasets 2 and 3; binomial test, $P = 0.008$), or 6 branches (dataset 1; binomial test, $P = 0.063$), and is significant for either two (datasets 1 and 3) or four branches (dataset 2). Furthermore, the correlation between pathway position and d_N is negative in all 7 branches for all datasets (binomial test, $P = 0.008$), and significant for either two (datasets 1 and 3) or six phylogenetic branches (dataset 2).

Effect of expression patterns, codon bias, protein length and connectivity on gene sequence evolution

We evaluated whether gene expression level and breadth, codon bias, protein length and connectivity correlate (1) with pathway position, (2) with the ω, d_N and d_S values, or (3) among them. We found that (tables 2, S8 and S9): (1) only protein length has a significant correlation with the pathway position, regardless of the dataset used ($\rho \leq -0.365$, $P \leq 0.026$). (2) d_S correlates significantly with ENC for datasets 1 and 2 ($\rho \leq -0.473$, $P \leq 0.030$); ω and d_N correlate with expression level for dataset 3 ($\rho \leq -0.498$, $P \leq 0.026$); and d_N correlates with protein length for dataset 2 ($\rho = 0.438$, $P = 0.047$). (3) Gene

expression breadth correlates with expression level for all datasets ($\rho \geq 0.606$, $P \leq 0.005$), and with the ENC ($\rho \geq 0.650$, $P \leq 0.002$) and connectivity ($\rho \geq 0.631$, $P \leq 0.003$) for dataset 3.

We also applied two multivariate analysis techniques (path analysis and partial correlation analysis) to evaluate the association between pathway position and ω, d_N and d_S after factoring out the discussed factors. Both analyses show that the association between pathway position and ω and d_N is always negative, regardless of the dataset used and both using 6 species or excluding *O. anatinus* (tables S12 and S13). The path analysis, furthermore, reveals a significant association between pathway position and d_N for dataset 1 (standardized path coefficient, $\beta = -0.246$; $P = 0.041$; table S12). Moreover, this analysis shows a significant association between pathway position and ω and d_N when *O. anatinus* is not considered (for dataset 3). The analyses conducted separately for each of the 9 phylogenetic branches show that the association between pathway position and both ω and d_N (but not d_S), is negative in a number of branches higher than the 50% expected by chance (tables 4, S14 and S15).

Connections of the IT pathway elements with other pathways

We studied the pattern of signaling interactions across the IT pathway proteins using the dataset reported by Cui et al. (2007). This manually-curated dataset consists of a directed graph with 1634 elements connected by 5089 interactions (2403 activatory, 741 inhibitory, 1915 undirected and 30 unspecified). Three hundred and fifty-six of these interactions (215 activatory, 74 inhibitory and 67 undirected; table S16) connect an IT with a non-IT pathway component. For each element, the number of inputs (received from other pathways) was computed as the number of arcs (interactions) that have the IT pathway protein as the head and the tail is not another IT pathway protein; conversely, the

number of outputs was computed as the number of interactions where the IT pathway is the tail and the head is not another IT pathway protein. In total, the IT pathway has 130 inputs (100 activatory and 30 inhibitory) and 159 outputs (115 activatory and 44 inhibitory; table S16).

DISCUSSION

We have characterized the evolutionary forces acting on the vertebrate IT pathway genes. All ω estimates are lower than 1 (with a maximum of 0.140; table S3), indicating that purifying selection is a major force acting on the IT pathway gene sequence evolution. This result, together with the fact that all genomes appear to encode at least one isoform of each IT pathway component, strongly supports that all studied organisms have a complete and functional IT pathway.

Polarity in selective constraint levels along the IT pathway

In *Drosophila*, we detected a correlation between the strength of purifying selection and position along the upstream/downstream axis of the IT pathway, being the downstream genes the most constrained (Alvarez-Ponce et al. 2009). Although in vertebrates this trend is not significant, the sign of the correlation coefficient is always negative, regardless of the metrics of selective constraint (ω or d_N) or the dataset used (tables 2, S8 and S9, figures 4 and S1). The analysis conducted separately for each phylogenetic branch (figure 5) reveals that the correlation coefficient is negative in a number of branches significantly greater than the number expected by chance (i.e., 50%), independently of the dataset used for d_N, and for datasets 2 and 3 for ω. This consistency in the direction of the association between selective constraint and pathway position across the vertebrate phylogeny is not compatible with a random distribution of selective constraint levels along the IT pathway. Furthermore, after removing the *O. anatinus* sequences from the analyses, the correlation between the pathway position and the overall d_N values is statistically significant and negative for dataset 2 (table 2).

Taken together, vertebrate results, as those in *Drosophila*, show a polarity in the levels of selective constraint along the IT pathway, with downstream elements evolving

under stronger purifying selection. This feature, therefore, would be neither incidental nor specific of the *Drosophila* genus, but it would rather result from a more general mechanism. This observation indicates that the molecular evolution pattern of the IT pathway components is affected by their particular position in the pathway. A correlation between pathway position and the strength of purifying selection has also been observed in other pathways, including the plant anthocyanin, isoprene, terpenoid, and carotenoid biosynthetic pathways (Rausher et al. 1999; Sharkey et al. 2005; Livingstone and Anderson 2009; Ramsay et al. 2009), and in the *Drosophila* Ras signal transduction pathway (Riley et al. 2003). However, the selective constraint polarity observed in these studies occurs in the opposite direction than in the IT pathway. Our results, therefore, indicate that this observation of higher selective constraint levels at the upstream part of molecular pathways is not universal.

The observed polarity of the selective constraints along the IT pathway might be generated by a putative polarity in a number of factors affecting evolutionary rate. For instance, if positive selection acted preferentially in the upstream part of the pathway, higher ω and d_N values would be expected at this part. However, we have identified the footprint of positive selection on only three genes (*IRS4*, *AKT3* and *PRKCD*, at pathway positions 1, 5 and 6, respectively), and none of the tests is significant after controlling for multiple testing (table S6). Positive selection, therefore, is not a major force shaping the evolution of the IT pathway genes, and should be discarded as an explanation for the ω and d_N polarity along the IT pathway.

Genes with higher expression levels and breadths, more biased codon usage, higher connectivities, or encoding shorter proteins tend to evolve under stronger purifying selection (Sharp 1991; Duret and Mouchiroud 2000; Pál et al. 2001; Fraser et al. 2002; Subramanian and Kumar 2004). A putative polarity in any of these factors, therefore,

might contribute to the observed selective constraint polarity along the pathway. Indeed, we have detected a negative correlation between protein length and pathway position, and d_N correlates positively with protein length for dataset 2. However, both partial correlation and path analysis show that the departure from 50% in the number of phylogenetic branches with negative sign in the association between pathway position and ω and d_N remains significant after controlling for the above factors (tables 4, S14, S15) and, therefore, they would not explain the detected correlation between selective constraint and pathway position.

Given that mutations in genes involved in a large number of pathways likely have important pleiotropic effects, these genes might be under strong selective constraint. Accordingly, in a linear pathway able to modulate the activation of other pathways (i.e., with signaling outputs along the pathway), upstream genes will be involved in a higher number of pathways and hence will evolve under stronger purifying selection. Conversely, a pathway that receives signaling inputs from other pathways is expected to be more constrained at the downstream part. The direction of the selective constraint polarity observed along the anthocyanin biosynthetic pathway (Rausher et al. 1999) would be consistent with this model since in this pathway upstream genes participate in the biosynthesis of a greater array of compounds than downstream genes, which are only involved in the biosynthesis of anthocyanins. The same reasoning would apply for the other biosynthetic pathways with a similar distribution of selective constraints (Sharkey et al. 2005; Livingstone and Anderson 2009; Ramsay et al. 2009).

Our results showing that downstream IT pathway genes evolve under stronger purifying selection than upstream genes might therefore be explained on the grounds of the IT pathway having more inputs than outputs. However, our analysis of the connection pattern of the IT pathway with other pathways shows that it has in fact more outputs than

inputs (table S16). Nevertheless, current knowledge of the IT pathway connection pattern is far from complete. Furthermore, given that signalling interactions differ in their biological impact, the number of inputs and outputs likely is a naive predictor of the distribution of selective constraint along the pathway. A more accurate predictor should include the relative biological significance of inputs and outputs in terms of fitness effects, although this is currently difficult to evaluate. Consequently, it is premature to draw conclusions about the effect of the connection pattern of the IT pathway on the evolution of its components.

Proteins in a pathway can contribute differentially to the overall pathway function. Enzymes with a high effect on the pathway function are expected to be under stronger natural selection than enzymes with a limited effect, which would be under weaker selection (Hartl et al. 1985; Eanes 1999; Watt and Dean 2000). Enzymes acting at network branch points are expected to play a key role in flux control, and hence to be preferentially targeted by natural selection (LaPorte et al. 1984; Stephanopoulos and Vallino 1991). Consistently, Flowers et. al. (2007) found that, in the *Drosophila* glucose catabolic pathways, positive selection preferentially targets genes acting on network branch points. Interestingly, two of the three IT pathway genes showing evidence of positive selection in vertebrates (*AKT3* and *PRKCD*) act in major network branch points. An analysis of the sensitivity of the IT pathway function to the kinetic properties of each of its components might provide insights into the distribution of purifying and positive selection along the pathway. The recent development of a mathematical model for the IT pathway (Zielinski et al. 2009) might serve as a starting point for this kind of analysis.

Physically interacting IT pathway proteins tend to evolve under similar selective constraint levels

We found that the levels of selective constraint of physically interacting proteins are more similar than those expected from a random network (Table S7). Such a tendency has also been observed in interactome-wide analyses (Fraser et al. 2002; Lemos et al. 2005), and has been explained by the putative coevolution among amino acids involved in PPIs, or by the similarity in the strength of stabilizing selection between interacting proteins (Fraser et al. 2002; Lemos et al. 2005). In the IT pathway, however, this pattern might be a by-product of the polarity in the selective constraint levels along the pathway. Since proteins tend to interact with those occupying adjacent positions in the pathway, the selective constraint polarity might determine that interacting proteins also exhibit similar selective constraint levels. Removing the effect of the association between pathway position and selective constraint, however, yielded equivalent results (table S17). The similarity in selective constraint levels, therefore, is independent of the polarity in selective constraint along the pathway, pointing to the molecular coevolution or some putative similarity in the strength of stabilizing selection among interacting proteins as the underlying biological mechanism.

These results in vertebrates contrast with our findings in *Drosophila* that the similarity in selective constraint levels among interacting IT pathway proteins vanished after controlling for the association between pathway position and selective constraint (Alvarez-Ponce et al. 2009). The number of interactions, however, remarkably differs between both studies (32 PPIs in vertebrates versus only 20 in *Drosophila*). Hence, the lack of significance in *Drosophila* might result from a lower statistical power associated with the smaller number of interactions. Consistently, when the analysis of the vertebrate IT pathway was restricted to the 20 interactions analyzed in *Drosophila*, we obtained

equivalent results in *Drosophila* and vertebrates: the association is significant for ω in dataset 2 (table S7), but not after controlling for the association between pathway position and selective constraint levels (table S17).

Molecular evolution of the *Drosophila* and vertebrate IT pathways

Although both *Drosophila* and vertebrates show a polarity in selective constraint levels along the IT pathway, the trend is less apparent in vertebrates. The difference might be explained by a lower statistical power of the vertebrate analysis caused by a putative smaller number of substitutions. However, the number of total synonymous changes (and the d_S values) are in fact higher in vertebrates than in *Drosophila* [paired *t* test, $P = 0.004$ for the number of synonymous changes; $P < 0.001$ for d_S (dataset 2)]. The lower effective population size of vertebrates (as compared with *Drosophila*; Lynch 2007) might also explain this difference. Indeed, the nearly neutral theory of molecular evolution predicts that natural selection will be more relaxed in populations with a small effective population size (Ohta 1973) and, in fact, purifying selection has been detected to be stronger in *Drosophila* than in mammals (e.g., Petit and Barbadilla 2009). Therefore, the putative biological mechanism maintaining the polarity in functional constraint levels along the IT pathway might be less efficient in vertebrates. However, there is no apparent reduction in the selective constraint levels among vertebrate genes, given that ω values do not differ significantly between vertebrates and *Drosophila* (paired *t* test, $P = 0.999$ for dataset 2).

Although in *Drosophila* most IT pathway genes are single-copy (Alvarez-Ponce et al. 2009), most pathway genes exist in multiple copies in vertebrates (table S4). Because the strength of purifying selection depends on the number of duplicates (Lynch and Conery 2000; Jordan et al. 2004), the selective constraint polarity observed along the IT pathway in vertebrates could result from a gradient in the number of duplicates. Nevertheless, since

the number of copies per paralogous group neither correlates with pathway position (ρ = –0.201, P = 0.383) nor with the average ω (ρ = –0.021, P = 0.923) or d_N values (ρ = 0.010, P = 0.963), this factor would not account for the selective constraint polarity either.

Concluding remarks

In summary, we provide evidence that the IT pathway architecture has an important effect on the patterns of molecular evolution of its components. We found a polarity in selective constraint levels along the vertebrate IT pathway, being downstream genes the most constrained. This selective constraint polarity mirrors that observed in *Drosophila* (Alvarez-Ponce et al. 2009). Therefore, although the biological mechanisms underlying this gradient distribution of selective constraints remain elusive, they are likely the same in *Drosophila* and vertebrates. The direction of the selective constraint polarity, however, differs from studies in a number of pathways showing that purifying selection is stronger in their upstream part (Rausher et al. 1999; Riley et al. 2003; Sharkey et al. 2005; Livingstone and Anderson 2009; Ramsay et al. 2009). Further understanding of the connection pattern of the IT pathway with other pathways, and of how its function depends on the properties of each of its elements, will probably provide insights into the factors underlying the pattern of molecular evolution of the IT pathway genes. Furthermore, the comprehensive study of pathways with different topologies will likely enhance our understanding of the effect of the pathway architecture on the molecular evolution of its components.

ACKNOWLEDGMENTS

This work was supported by the Ministerio de Educación y Ciencia (Spain) [BFU2007-62927 to J.R. and BFU2007-63228 to M.A.] and the Comissió Interdepartamental de Recerca i Innovació Tecnològica (Spain) [2005SGR-00166 and 2009SGR-1287]. D.A-P. was supported by a predoctoral fellowship from the Ministerio de Educación y Ciencia (Spain) [AP2005-0012].

REFERENCES

Alvarez-Ponce D, Aguadé M, Rozas J. 2009. Network-level molecular evolutionary analysis of the insulin/TOR signal transduction pathway across 12 Drosophila genomes. Genome Res. 19:234-242.

Antoine M, Fried M. 1992. The organization of the intron-containing human S6 ribosomal protein (rpS6) gene and determination of its location at chromosome 9p21. Hum Mol Genet. 1:565-570.

Benjamini Y, Hochberg Y. 1995. Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing. J. R. Statist. Soc. B. 57:289-300.

Bossi A, Lehner B. 2009. Tissue specificity and the human protein interaction network. Mol Syst Biol. 5:260.

Castillo-Davis CI, Kondrashov FA, Hartl DL, Kulathinal RJ. 2004. The functional genomic distribution of protein divergence in two animal phyla: coevolution, genomic conflict, and constraint. Genome Res. 14:802-811.

Cork JM, Purugganan MD. 2004. The evolution of molecular genetic pathways and networks. Bioessays. 26:479-484.

Cui Q et al. 2007. A map of human cancer signaling. Mol Syst Biol. 3:152.

Do CB, Mahabhashyam MS, Brudno M, Batzoglou S. 2005. ProbCons: Probabilistic consistency-based multiple sequence alignment. Genome Res. 15:330-340.

Duret L, Mouchiroud D. 2000. Determinants of substitution rates in mammalian genes: expression pattern affects selection intensity but not mutation rate. Mol Biol Evol. 17:68-74.

Eanes WF. 1999. Analysis of selection on enzyme polymorphisms. Rev Ecol Syst. 30:301-326.

Elsik CG et al. 2009. The genome sequence of taurine cattle: a window to ruminant biology and evolution. Science. 324:522-528.

Feo S, Davies B, Fried M. 1992. The mapping of seven intron-containing ribosomal protein genes shows they are unlinked in the human genome. Genomics. 13:201-207.

Flicek P et al. 2008. Ensembl 2008. Nucleic Acids Res. 36:D707-714.

Flowers JM et al. 2007. Adaptive evolution of metabolic pathways in Drosophila. Mol Biol Evol. 24:1347-1354.

Fraser HB, Hirsh AE, Steinmetz LM, Scharfe C, Feldman MW. 2002. Evolutionary rate in the protein interaction network. Science. 296:750-752.

Fryxell KJ. 1996. The coevolution of gene family trees. Trends Genet. 12:364-369.

Goldman N, Yang Z. 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol Biol Evol. 11:725-736.

Hahn MW, Kern AD. 2005. Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks. Mol Biol Evol. 22:803-806.

Hall TA. 1999. BioEdit: A user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucl Acids Symp Ser. 41:95-98.

Hartl DL, Dykhuizen DE, Dean AM. 1985. Limits of adaptation: the evolution of selective neutrality. Genetics. 111:655-674.

Ihaka R, Gentleman R. 1996. R: a language for data analysis and graphics. J Comp Graph Stat. 5:299-314.

Ingvarsson PK. 2007. Gene expression and protein length influence codon usage and rates of sequence evolution in Populus tremula. Mol Biol Evol. 24:836-844.

International Chicken Genome Sequencing Consortium. 2004. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. Nature. 432:695-716.

International Human Genome Sequencing Consortium. 2004. Finishing the euchromatic sequence of the human genome. Nature. 431:931-945.

Jones DT, Taylor WR, Thornton JM. 1992. The rapid generation of mutation data matrices from protein sequences. Comput Appl Biosci. 8:275-282.

Jordan IK, Wolf YI, Koonin EV. 2004. Duplicated genes evolve slower than singletons despite the initial rate increase. BMC Evol Biol. 4:22.

Jovelin R, Dunham JP, Sung FS, Phillips PC. 2009. High nucleotide divergence in developmental regulatory genes contrasts with the structural elements of olfactory pathways in caenorhabditis. Genetics. 181:1387-1397.

Kacser H, Burns JA. 1973. The control of flux. Symp Soc Exp Biol. 27:65-104.

Kimura M. 1983. The neutral theory of molecular evolution. Cambridge University Press, Cambridge

LaPorte DC, Walsh K, Koshland DE, Jr. 1984. The branch point effect. Ultrasensitivity and subsensitivity to metabolic control. J Biol Chem. 259:14068-14075.

Lemos B, Bettencourt BR, Meiklejohn CD, Hartl DL. 2005. Evolution of proteins and gene expression levels are coupled in Drosophila and are independently associated with mRNA abundance, protein length, and number of protein-protein interactions. Mol Biol Evol. 22:1345-1354.

LeRoith D, Taylor SI, Olefsky JM. 2004. Diabetes mellitus: a fundamental and clinical text. Lippincott Williams & Wilkins, Philadelphia.

Librado P, Rozas J. 2009. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. Bioinformatics. 25:1451-1452.

Livingstone K, Anderson S. 2009. Patterns of variation in the evolution of carotenoid biosynthetic pathway enzymes of higher plants. J Hered. 100:754-761.

Lu Y, Rausher MD. 2003. Evolutionary rate variation in anthocyanin pathway genes. Mol Biol Evol. 20:1844-1853.

Lynch M. 2007. The origins of genome architecture. Sinauer Associates, Sunderland, Mass.

Lynch M, Conery JS. 2000. The evolutionary fate and consequences of duplicate genes. Science. 290:1151-1155.

Mikkelsen TS et al. 2007. Genome of the marsupial Monodelphis domestica reveals innovation in non-coding sequences. Nature. 447:167-177.

Mouse Genome Sequencing Consortium. 2002. Initial sequencing and comparative analysis of the mouse genome. Nature. 420:520-562.

Ohta T. 1973. Slightly deleterious mutant substitutions in evolution. Nature. 246:96-98.

Oldham S, Hafen E. 2003. Insulin/IGF and target of rapamycin signaling: a TOR de force in growth control. Trends Cell Biol. 13:79-85.

Olsen KM, Womack A, Garrett AR, Suddith JI, Purugganan MD. 2002. Contrasting evolutionary forces in the Arabidopsis thaliana floral developmental pathway. Genetics. 160:1641-1650.

Pál C, Papp B, Hurst LD. 2001. Highly expressed genes in yeast evolve slowly. Genetics. 158:927-931.

Pata I, Metspalu A. 1996. Structural characterization of the mouse ribosomal protein S6-encoding gene. Gene. 175:241-245.

Petit N, Barbadilla A. 2009. The efficiency of purifying selection in Mammals vs. Drosophila for metabolic genes. J Evol Biol. 22:2118-2124.

Ponting CP. 2008. The functional repertoires of metazoan genomes. Nat Rev Genet. 9:689-698.

Puig O, Tjian R. 2005. Transcriptional feedback control of insulin receptor by dFOXO/FOXO1. Genes Dev. 19:2435-2446.

Ramsay H, Rieseberg LH, Ritland K. 2009. The correlation of evolutionary rate with pathway position in plant terpenoid biosynthesis. Mol Biol Evol. 26:1045-1053.

Rat Genome Sequencing Project Consortium. 2004. Genome sequence of the Brown Norway rat yields insights into mammalian evolution. Nature. 428:493-521.

Rausher MD, Lu Y, Meyer K. 2008. Variation in constraint versus positive selection as an explanation for evolutionary rate variation among anthocyanin genes. J Mol Evol. 67:137-144.

Rausher MD, Miller RE, Tiffin P. 1999. Patterns of evolutionary rate variation among genes of the anthocyanin biosynthetic pathway. Mol Biol Evol. 16:266-274.

Riley RM, Jin W, Gibson G. 2003. Contrasting selection pressures on components of the Ras-mediated signal transduction pathway in Drosophila. Mol Ecol. 12:1315-1323.

Sharkey TD et al. 2005. Evolution of the isoprene biosynthetic pathway in kudzu. Plant Physiol. 137:700-712.

Sharp PM. 1991. Determinants of DNA sequence divergence between Escherichia coli and Salmonella typhimurium: codon usage, map position, and concerted evolution. J Mol Evol. 33:23-33.

Stephanopoulos G, Vallino JJ. 1991. Network rigidity and metabolic engineering in metabolite overproduction. Science. 252:1675-1681.

Su AI et al. 2002. Large-scale analysis of the human and mouse transcriptomes. Proc Natl Acad Sci U S A. 99:4465-4470.

Su AI et al. 2004. A gene atlas of the mouse and human protein-encoding transcriptomes. Proc Natl Acad Sci U S A. 101:6062-6067.

Subramanian S, Kumar S. 2004. Gene expression intensity shapes evolutionary rates of the proteins encoded by the vertebrate genome. Genetics. 168:373-381.

Taguchi A, White MF. 2008. Insulin-like signaling, nutrient homeostasis, and life span. Annu Rev Physiol. 70:191-212.

Tamura K, Dudley J, Nei M, Kumar S. 2007. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. Mol Biol Evol. 24:1596-1599.

Tamura K, Nei M. 1993. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Mol Biol Evol. 10:512-526.

Torrents D, Suyama M, Zdobnov E, Bork P. 2003. A genome-wide survey of human pseudogenes. Genome Res. 13:2559-2567.

Vinciguerra M, Foti M. 2006. PTEN and SHIP2 phosphoinositide phosphatases as negative regulators of insulin signalling. Arch Physiol Biochem. 112:89-104.

Wang D, Hsieh M, Li WH. 2005. A general tendency for conservation of protein length across eukaryotic kingdoms. Mol Biol Evol. 22:142-147.

Warren WC et al. 2008. Genome analysis of the platypus reveals unique signatures of evolution. Nature. 453:175-183.

Watt WB, Dean AM. 2000. Molecular-functional studies of adaptive genetic variation in prokaryotes and eukaryotes. Annu Rev Genet. 34:593-622.

Whelan S, Goldman N. 1999. Distributions of Statistics Used for the Comparison of Models of Sequence Evolution in Phylogenetics. Mol Biol Evol. 16:1292-1299.

Wong WS, Yang Z, Goldman N, Nielsen R. 2004. Accuracy and power of statistical methods for detecting adaptive evolution in protein coding sequences and for identifying positively selected sites. Genetics. 168:1041-1051.

Wright F. 1990. The 'effective number of codons' used in a gene. Gene. 87:23-29.

Yang YH, Zhang FM, Ge S. 2009. Evolutionary rate patterns of the Gibberellin pathway genes. BMC Evol Biol. 9:206.

Yang Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci. 13:555-556.

Yang Z, Nielsen R, Goldman N, Pedersen AM. 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics. 155:431-449.

Yang Z, Wong WS, Nielsen R. 2005. Bayes empirical bayes inference of amino acid sites under positive selection. Mol Biol Evol. 22:1107-1118.

Zhang Z, Harrison P, Gerstein M. 2002. Identification and analysis of over 2000 ribosomal protein pseudogenes in the human genome. Genome Res. 12:1466-1482.

Zielinski R et al. 2009. The crosstalk between EGF, IGF, and Insulin cell signaling pathways--computational and experimental analysis. BMC Syst Biol. 3:88.

FIGURES

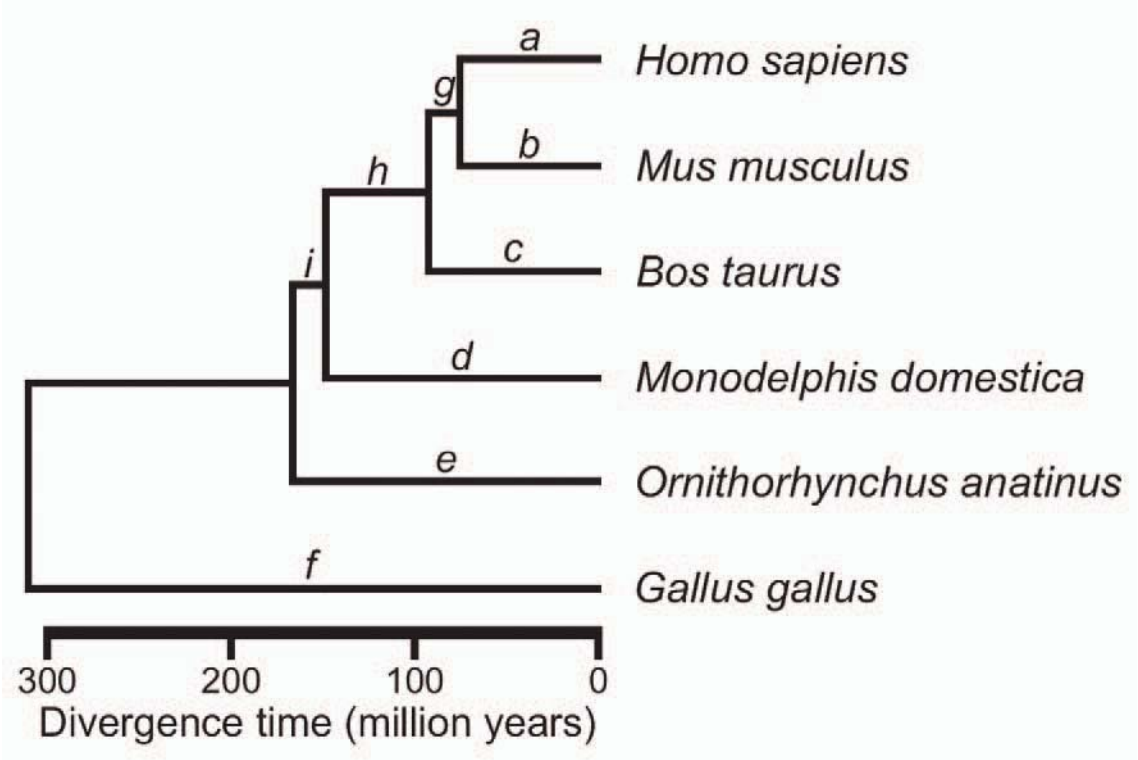

Fig. 1. Phylogenetic relationships among the six vertebrate species analyzed. Tree topology and divergence times were taken from Ponting (2008).

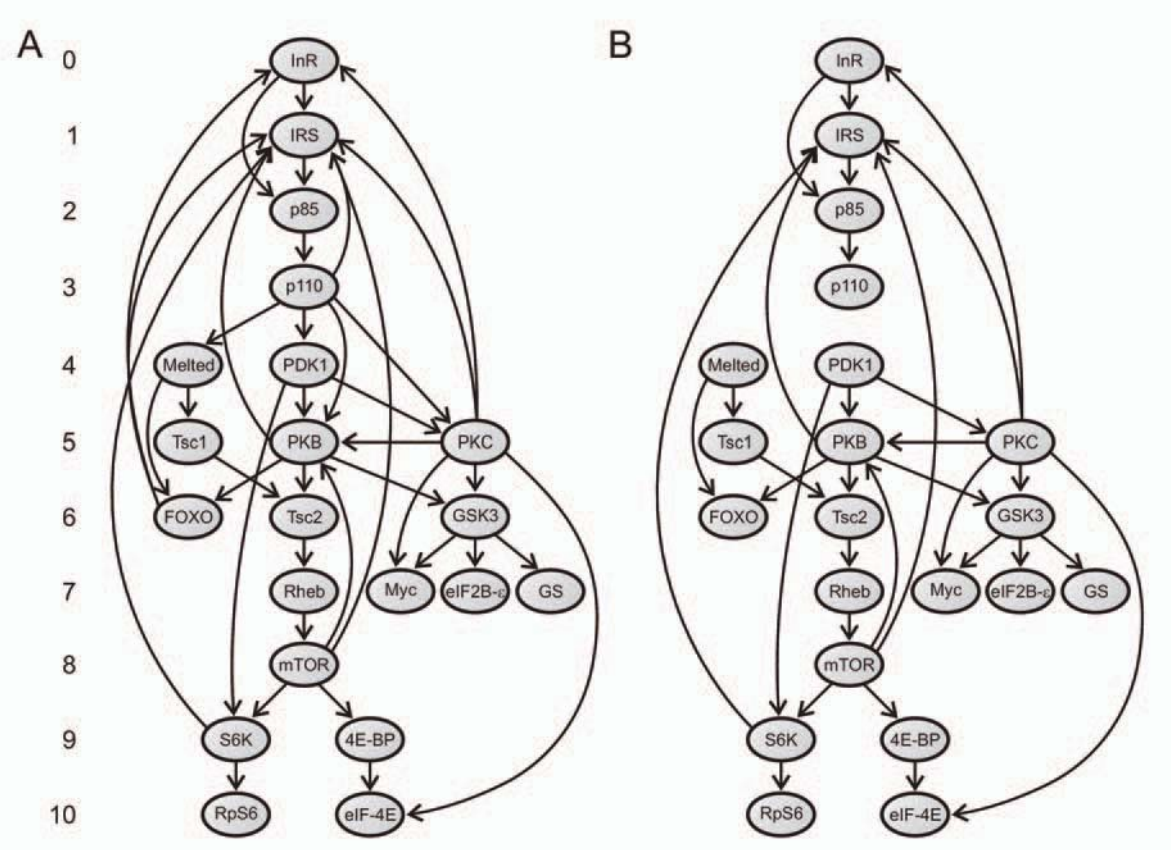

Fig. 2. Directed graphs used in the network-level analyses. (A) Graph *G* containing all interactions (arcs) among human IT pathway proteins (nodes). This graph consists of 21 nodes and 39 arcs, of which 32 represent physical protein–protein interactions (PPIs), 5 involve the membrane phospholipid PIP_3 (synthesized by p110 isoforms and activates the IRS, Melted, PDK1, PKB and PKC proteins), and the other 2 represent the activation of *INR* and *IRS2* genes by the FOXO transcription factors (Puig and Tjian 2005). Numbers on the left part represent the position of the elements in the pathway. Human proteins with orthologs in a previous *Drosophila* IT pathway study (Alvarez-Ponce et al. 2009) were assigned the same position than their *Drosophila* counterparts. We assigned the position 5 to PKC proteins, as they are activated by PDK1 (position 4) (LeRoith et al. 2004). We excluded the phosphoinositide phosphatase PTEN from network-level analyses, because it does not directly interact with any other element in the graph (see Vinciguerra and Foti 2006 for review). The cytohesins Cyh1–4 were also excluded as their specific function in the pathway remains unclear (Hafner et al. 2006). (B) Graph *S*, subgraph of *G* containing only the 32 physical interactions.

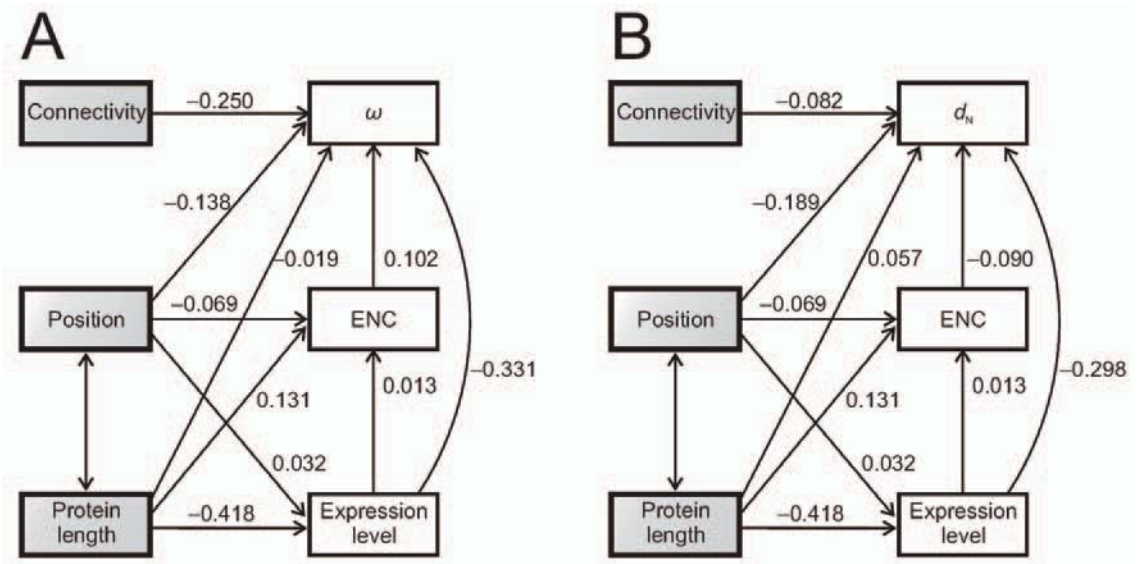

Fig. 3. Path analysis for dataset 2. Single- and double-headed arrows represent assumed causal dependencies and correlations, respectively. Numbers in each arrow represent the standardized path coefficients (β). None of the associations was significant. The analyses conducted using expression breadth instead of expression level yielded equivalent results. (A) Analysis for ω. (B) Analysis for d_N.

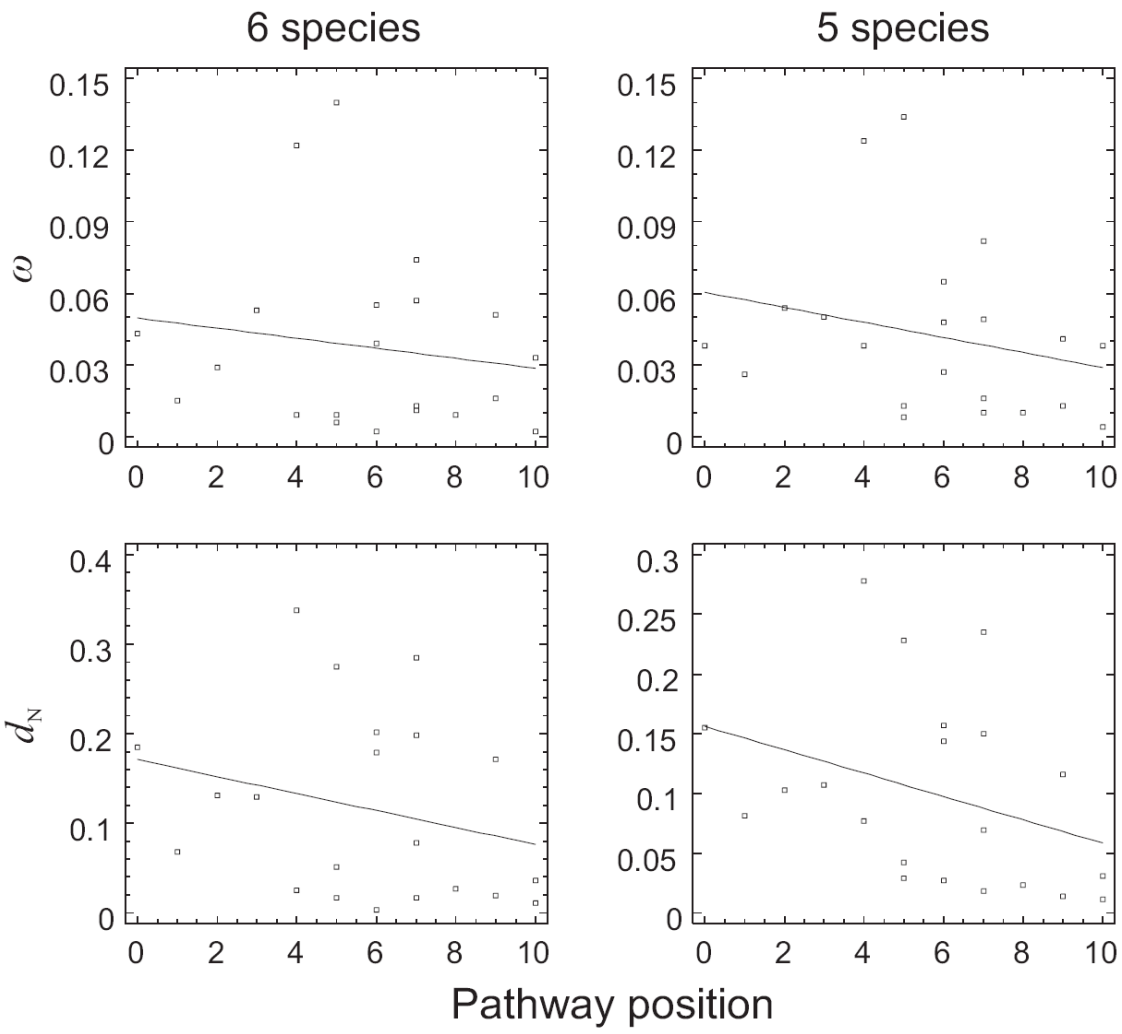

Fig. 4. Correlation between pathway position and ω and d_N (dataset 2), using all six species, or excluding *O. anatinus*. Continuous lines represent regression lines. An extended version of this figure is provided as supplementary material (figure S1).

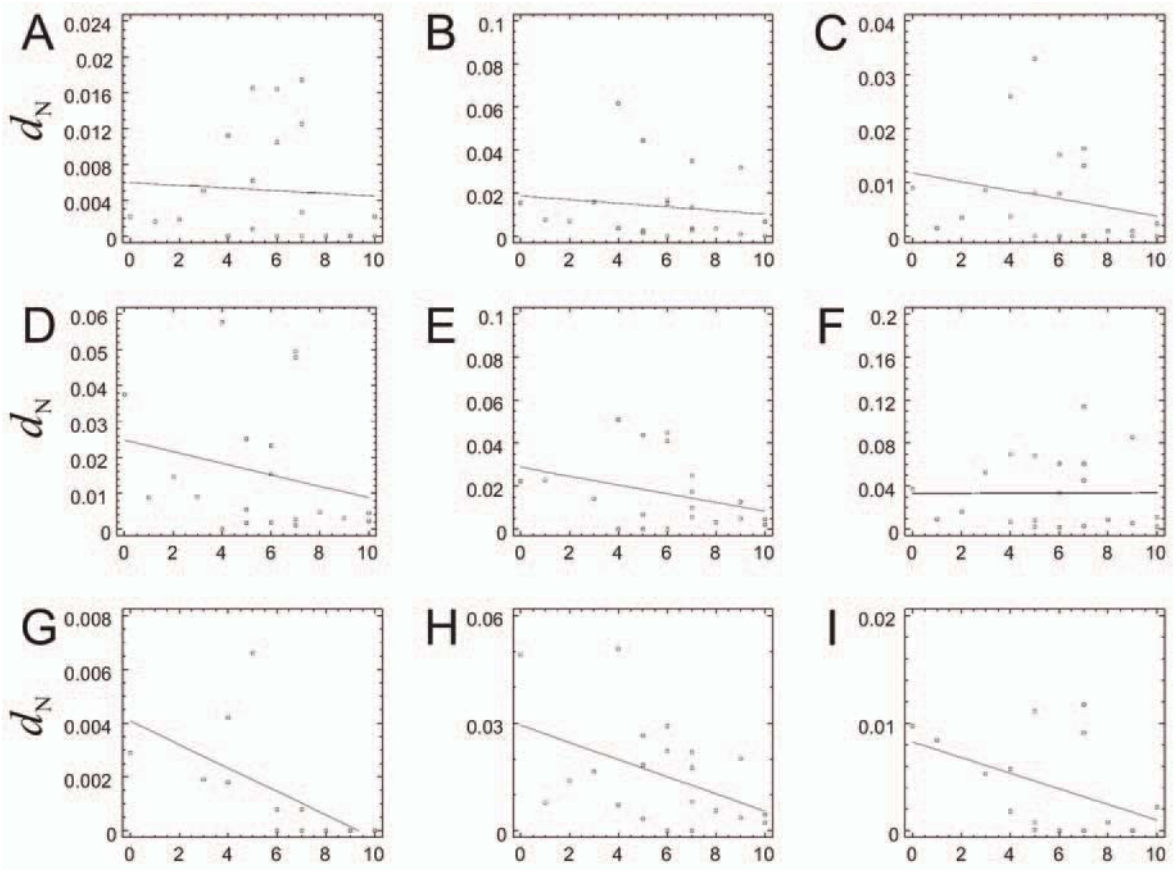

Fig. 5. Correlation between pathway position and d_N in all 9 phylogenetic branches (dataset 2). Panels A–I correspond to branches a–i in figure 1. Continuous lines represent regression lines.

Table 1. Dataset 2

Gene	Pathway position	6 species					5 species[a]					Gene expression		Connectivity	Protein length[f]
		d_N[b]	d_S[b]	ω	ENC	%used codons[c]	d_N[b]	d_S[b]	ω	ENC	%used codons[c]	Level[d]	Breadth[e]		
INSR	0	0.185	4.344	0.043	49.77	77.50	0.155	4.131	0.038	49.52	91.61	1012.74	22	73	1382
IRS1	1	0.068	4.688	0.015	41.98	23.19	0.081	3.096	0.026	42.95	48.07	340.74	18	66	1242
PIK3R1	2	0.131	4.610	0.029	55.14	77.73	0.103	1.912	0.054	55.64	98.63	835.66	25	132	732
PIK3CB	3	0.129	2.430	0.053	52.82	95.05	0.107	2.146	0.050	52.71	99.25	419.30	24	7	1070
VEPH1	4	0.338	2.774	0.122	53.15	92.80	0.278	2.235	0.124	52.62	92.80	80.54	2	4	833
PDPK1	4	0.025	2.916	0.009	52.72	44.06	0.077	2.064	0.038	52.54	80.04	1338.84	25	36	556
AKT1	5	0.051	6.026	0.009	45.77	74.38	0.042	4.984	0.008	40.99	92.71	970.02	19	108	480
PRKCI	5	0.017	2.722	0.006	55.98	88.09	0.029	2.208	0.013	56.39	95.13	1147.88	25	33	596
TSC1	5	0.275	1.970	0.140	54.69	91.49	0.228	1.708	0.134	54.54	91.49	715.78	25	15	1164
FOXO1	6	0.202	3.688	0.055	47.50	74.35	0.157	2.419	0.065	49.32	87.63	868.36	25	23	655
GSK3B	6	0.003	1.551	0.002	53.88	62.59	0.027	0.974	0.027	53.99	99.77	561.56	25	88	433
TSC2	6	0.179	4.578	0.039	45.64	83.45	0.144	3.032	0.048	49.98	94.08	358.20	17	22	1807
EIF2B5	7	0.285	3.861	0.074	53.98	86.82	0.235	2.876	0.082	54.04	91.40	711.14	25	68	721
GYS1	7	0.078	5.856	0.013	41.29	50.07	0.069	7.092	0.010	42.24	50.07	1352.30	25	10	737
MYC	7	0.198	3.513	0.057	43.58	47.58	0.150	3.056	0.049	44.20	73.57	490.34	18	148	454
RHEB	7	0.017	1.481	0.011	45.33	90.22	0.018	1.115	0.016	46.09	100.00	2303.66	25	7	184
MTOR	8	0.027	2.916	0.009	49.47	94.00	0.023	2.374	0.010	50.68	96.43	302.60	20	15	2549
EIF4EBP1	9	0.171	3.370	0.051	43.46	38.14	0.116	2.849	0.041	45.02	68.64	558.70	19	15	118
RPS6KB1	9	0.019	1.188	0.016	53.83	89.90	0.014	1.020	0.013	53.44	94.48	164.44	7	21	525
EIF4E	10	0.036	1.081	0.033	54.81	94.93	0.031	0.815	0.038	54.20	94.93	—	—	106	217
RPS6	10	0.011	4.795	0.002	50.80	81.53	0.011	2.570	0.004	51.18	99.20	23490.48	25	179	249

ENC, effective number of codons (Wright 1990).
[a]Excluding *O. anatinus* from the analyses.
[b]Values estimated as the sum across all branches of the phylogeny.
[c]Percent of used codons for estimating d_N, d_S and ω values.
[d]Expression levels averaged across 25 selected human tissues (table S2).
[e]Number of selected tissues with expression level ≥ 200.
[f]Number of amino acids of the human protein.

Table 2. Bivariate correlations (dataset 2).

	Position		ω		d_N		d_S		ENC		%used codons		Expression level		Exp. breadth		Connectivity		Protein length	
	ρ	P	ρ	P	ρ	P	ρ	P	ρ	P	ρ	P	ρ	P	ρ	P	ρ	P	ρ	P
Position	—	—	-0.304	0.090	-0.441	0.023 *	-0.164	0.477	-0.050	0.828	0.110	0.636	0.032	0.894	0.011	0.962	0.040	0.862	-0.553	0.009 **
ω	-0.136	0.279	—	—	0.844	<0.001 ***	-0.190	0.410	0.353	0.116	-0.183	0.427	-0.355	0.125	-0.029	0.905	-0.173	0.453	0.230	0.316
d_N	-0.294	0.098	0.914	<0.001 ***	—	—	0.325	0.151	-0.005	0.982	-0.522	0.015 *	-0.277	0.238	-0.156	0.511	-0.150	0.517	0.413	0.063
d_S	-0.229	0.318	-0.105	0.650	0.229	0.319	—	—	-0.714	<0.001 ***	-0.610	0.003 **	0.087	0.715	-0.284	0.225	0.129	0.578	0.322	0.154
ENC	-0.107	0.646	0.027	0.907	-0.068	0.771	-0.473	0.030 *	—	—	0.487	0.025 *	-0.096	0.686	0.340	0.142	0.062	0.788	0.039	0.867
%used codons	0.144	0.532	0.200	0.385	0.004	0.987	-0.558	0.009 **	0.542	0.011 *	—	—	0.060	0.801	0.198	0.403	0.003	0.991	-0.156	0.500
Expression level	0.032	0.894	-0.423	0.063	-0.337	0.146	0.280	0.232	-0.026	0.915	-0.232	0.326	—	—	0.762	<0.001 ***	0.273	0.243	-0.389	0.090
Expression breadth	0.011	0.962	-0.274	0.242	-0.252	0.283	-0.014	0.952	0.337	0.146	-0.031	0.896	0.762	<0.001 ***	—	—	0.161	0.498	-0.224	0.342
Connectivity	0.040	0.862	-0.291	0.201	-0.175	0.448	0.367	0.101	0.183	0.426	-0.369	0.100	0.273	0.243	0.161	0.498	—	—	-0.291	0.200
Protein length	-0.553	0.009 *	0.292	0.199	0.438	0.047 *	0.261	0.253	-0.014	0.951	0.195	0.397	-0.389	0.090	-0.224	0.342	-0.291	0.200	—	—

*, $P < 0.05$; **, $P < 0.01$; ***, $P < 0.001$. Values below and above the principal diagonal are based on analyses using all six species or excluding *O. anatinus* from the analyses, respectively. All correlations are based on $n = 21$ observations except those involving gene expression level and breadth ($n = 20$). Two-tailed P-values are provided except for the correlations between pathway position and ω and d_N (one-tailed).

Table 3. Correlations between pathway position and ω, d_N and d_S for each phylogenetic branch (dataset 2).

			ω		d_N		d_S	
# species	Branch[a]	*n*	ρ	P[b]	ρ	P[b]	ρ	P[c]
6	all[d]	21	-0.136	0.279	-0.294	0.098	-0.229	0.318
	a	21	-0.252	0.136	-0.255	0.133	-0.108	0.642
	b	21	-0.116	0.309	-0.314	0.083	-0.505	0.020 *
	c	21	-0.382	0.044 *	-0.422	0.028 *	-0.409	0.065
	d	20	-0.115	0.315	-0.278	0.118	-0.190	0.422
	e	20	-0.198	0.201	-0.331	0.077	-0.183	0.439
	f	21	-0.101	0.332	-0.096	0.339	-0.190	0.409
	g	13	-0.828	<0.001 ***	-0.824	0.001 ***	0.030	0.922
	h	21	-0.316	0.082	-0.394	0.039 *	-0.144	0.534
	i	16	-0.392	0.067	-0.439	0.045 *	0.071	0.794
5[e]	all[d]	21	-0.304	0.090	-0.441	0.023 *	-0.164	0.477
	a	21	-0.434	0.025 *	-0.455	0.019 *	-0.254	0.267
	b	21	-0.279	0.111	-0.395	0.038 *	-0.408	0.066
	c	21	-0.500	0.011 *	-0.482	0.014 *	-0.389	0.081
	d	20	-0.339	0.072	-0.426	0.031 *	-0.080	0.739
	f+i	21	-0.196	0.197	-0.214	0.176	-0.193	0.402
	g	13	-0.797	0.001 ***	-0.766	0.001 **	0.011	0.971
	h	21	-0.446	0.022 *	-0.549	0.005 *	-0.307	0.175

*, $P < 0.05$; **, $P < 0.01$; ***, $P < 0.001$.
[a]Branch codes according to figure 1.
[b]One-tailed *P*-values.
[c]Two-tailed *P*-values.
[d]Using overall ω, d_N and d_S values.
[e]Excluding *O. anatinus* from the analysis.

Table 4. Partial correlation and path analysis (dataset 2)

			Partial correlation analysis						Path analysis					
			ω		d_N		d_S		ω		d_N		d_S	
# species	Branch[a]	*n*	ρ	P[b]	ρ	P[b]	ρ	P[c]	β	P[b]	β	P[b]	β	P[c]
6	all[d]	20	-0.144	0.299	-0.117	0.335	0.084	0.762	-0.138	0.276	-0.189	0.213	-0.156	0.380
	a	20	-0.252	0.174	-0.173	0.264	0.457	0.064	-0.121	0.313	-0.076	0.379	-0.098	0.486
	b	20	-0.172	0.264	-0.210	0.219	-0.399	0.117	-0.102	0.330	-0.208	0.179	-0.276	0.122
	c	20	-0.121	0.330	-0.163	0.276	-0.118	0.667	-0.049	0.414	-0.071	0.379	-0.276	0.127
	d	19	-0.203	0.236	-0.102	0.361	0.141	0.622	-0.134	0.284	-0.106	0.317	0.047	0.834
	e	19	-0.141	0.310	-0.098	0.367	0.103	0.720	-0.126	0.301	-0.107	0.322	0.142	0.483
	f	20	-0.081	0.385	0.014	0.520	0.197	0.470	-0.015	0.476	0.003	0.505	0.020	0.934
	g	13	-0.817	<0.001 ***	-0.812	<0.001 ***	-0.124	0.760	-0.587	0.004 **	-0.649	<0.001 ***	-0.265	0.370
	h	20	-0.190	0.243	-0.173	0.263	0.226	0.404	-0.184	0.200	-0.302	0.092	-0.154	0.487
	i	15	-0.025	0.472	-0.164	0.319	-0.007	0.984	-0.189	0.079	-0.381	0.041 *	0.012	0.963
5[e]	all[d]	20	-0.408	0.054	-0.321	0.111	0.236	0.381	-0.307	0.074	-0.307	0.090	-0.033	0.847
	a	20	-0.449	0.035 *	-0.349	0.090	0.299	0.259	-0.315	0.085	-0.294	0.097	-0.085	0.566
	b	20	-0.393	0.062	-0.339	0.097	-0.085	0.759	-0.277	0.100	-0.366	0.042 *	-0.255	0.144
	c	20	-0.414	0.051	-0.273	0.153	-0.121	0.659	-0.197	0.169	-0.214	0.164	-0.177	0.338
	d	19	-0.381	0.077	-0.256	0.180	0.242	0.388	-0.369	0.046 *	-0.276	0.102	0.051	0.819
	f+i	20	-0.309	0.121	-0.145	0.298	0.025	0.929	-0.218	0.171	-0.135	0.293	-0.003	0.987
	g	13	-0.721	0.005 **	-0.739	0.004 **	-0.081	0.843	-0.561	0.018 *	-0.736	0.001 **	-0.164	0.492
	h	20	-0.263	0.163	-0.346	0.092	-0.097	0.726	-0.221	0.150	-0.358	0.039 *	-0.165	0.407

Association between pathway position and ω, d_N and d_S values after controlling for expression level and breadth, codon bias, protein length and connectivity. *, $P < 0.05$; **, $P < 0.01$; ***, $P < 0.001$.

[a]Branch codes according to figure 1.

[b]One-tailed *P*-values.

[c]Two-tailed *P*-values.

[d]Using overall ω, d_N and d_S values.

[e]Excluding *O. anatinus* from the analysis.

SUPPLEMENTARY MATERIAL

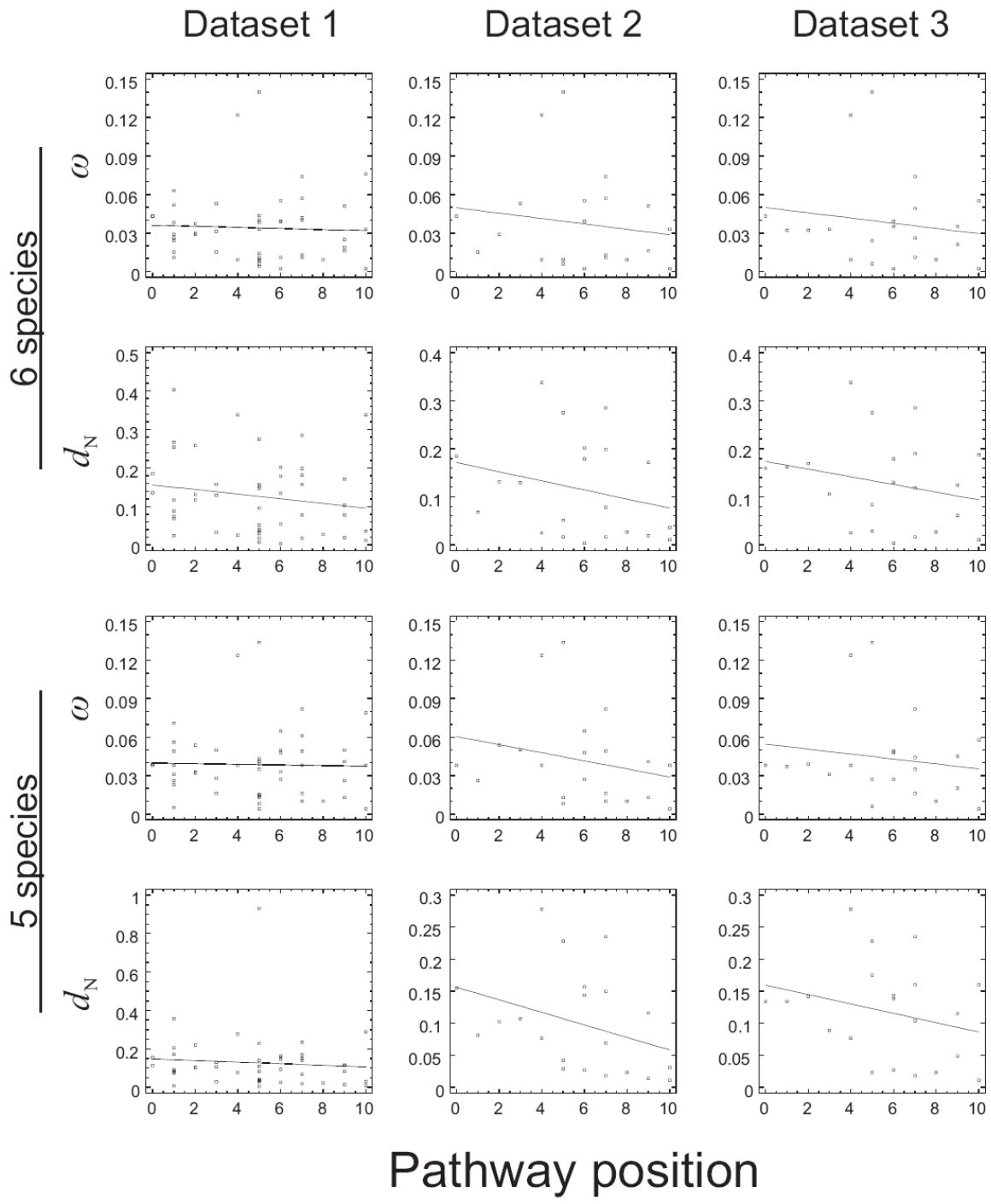

Fig. S1. Correlation between pathway position and ω and d_N, using all six species, or excluding *O. anatinus*. Continuous lines represent regression lines.

Table S1. List of genes involved in the insulin/TOR signal transduction pathway

								Genomic location			
Group	Accession[a]	Symbol[b]	Pseudogene[c]	Protein	#proteins[d]	Chosen isoform	#introns[e]	Chr	Strand	Start	End
INR	ENSG00000171105	*INSR*	—	InR	2	ENST00000302850	21	19	–	7063266	7245011
INR	ENSG00000140443	*IGF1R*	—	IGF1R	1	ENST00000268035	20	15	+	97010302	97325282
INR	ENSG00000027644	*INSRR*	—	IRR	1	ENST00000368195	21	1	–	155077289	155095290
IRS	ENSG00000169047	*IRS1*	—	IRS1	1	ENST00000305123	0	2	–	227308182	227372719
IRS	ENSG00000185950	*IRS2*	—	IRS2	1	ENST00000375856	1	13	–	109204185	109236916
IRS	ENSG00000184414	*AC069281.1*	yes	IRS3	1	ENST00000223076	0	7	+	100002657	100005961
IRS	ENSG00000133124	*IRS4*	—	IRS4	1	ENST00000372129	0	X	–	107862368	107866295
IRS	ENSG00000115325	*DOK1*	—	DOK1	2	ENST00000233668	4	2	+	74635355	74638181
IRS	Pseudogene:10	—	yes	—	—	—	0	11	–	71638542	71639980
IRS	ENSG00000147443	*DOK2*	—	DOK2	1	ENST00000276420	4	8	–	21822336	21827151
IRS	ENSG00000146094	*DOK3*	—	DOK3	3	ENST00000357198	5	5	–	176863358	176869459
IRS	ENSG00000125170	*DOK4*	—	DOK4	1	ENST00000340099	7	16	–	56063381	56077828
IRS	ENSG00000101134	*DOK5*	—	DOK5	2	ENST00000262593	7	20	+	52525570	52701097
IRS	ENSG00000206052	*DOK6*	—	DOK6	1	ENST00000382713	7	18	+	65219271	65660357
IRS	ENSG00000166225	*FRS2*	—	FRS2	1	ENST00000299293	4	12	+	68150396	68259826
IRS	ENSG00000137218	*FRS3*	—	FRS3	1	ENST00000259748	4	6	–	41845892	41853897
P85	ENSG00000145675	*PIK3R1*	—	p85α	4	ENST00000396611	14	5	+	67558218	67633403
P85	ENSG00000105647	*PIK3R2*	—	p85β	1	ENST00000222254	14	19	+	18125016	18142343
P85	ENSG00000117461	*PIK3R3*	—	p55γ	3	ENST00000372006	9	1	–	46278399	46371053
P110	ENSG00000121879	*PIK3CA*	—	p110α	1	ENST00000263967	19	3	+	180349005	180435189
P110	ENSG00000051382	*PIK3CB*	—	p110β	1	ENST00000289153	21	3	–	139856921	139960875
P110	ENSG00000105851	*PIK3CG*	—	p110γ	1	ENST00000359195	9	7	+	106293160	106334801
P110	ENSG00000171608	*PIK3CD*	—	p110δ	3	ENST00000377346	21	1	+	9634377	9711759
MELT	ENSG00000197415	*VEPH1*	—	Melt	2	ENST00000362010	12	3	–	158460228	158703830
PDK1	ENSG00000140992	*PDPK1*	—	PDK1	4	ENST00000342085	13	16	+	2527971	2593189
PDK1	ENSG00000205918	*AC141586.2*	—	—	1	ENST00000382326	8	16	–	2606123	2633298
AKT	ENSG00000142208	*AKT1*	—	PKBα	1	ENST00000402615	12	14	–	104306734	104330983
AKT	ENSG00000105221	*AKT2*	—	PKBβ	4	ENST00000392038	12	19	–	45428064	45483105
AKT	ENSG00000117020	*AKT3*	—	PKBγ	4	ENST00000366539	12	1	–	241731688	242080053
PKC	ENSG00000154229	*PRKCA*	—	PKCα	1	ENST00000284384	16	17	+	61729216	62237324
PKC	ENSG00000166501	*PRKCB*	—	PKCβ	2	ENST00000303531	16	16	+	23754823	24134803
PKC	ENSG00000126583	*PRKCG*	—	PKCγ	1	ENST00000263431	17	19	+	59077279	59102713
PKC	ENSG00000163932	*PRKCD*	—	PKCδ	2	ENST00000394729	16	3	+	53170263	53201771
PKC	ENSG00000171132	*PRKCE*	—	PKCε	2	ENST00000306156	14	2	+	45732547	46268632
PKC	ENSG00000067606	*PRKCZ*	—	PKCζ	3	ENST00000378567	17	1	+	1971769	2106692
PKC	ENSG00000027075	*PRKCH*	—	PKCη	1	ENST00000332981	13	14	+	60858186	61087443
PKC	ENSG00000065675	*PRKCQ*	—	PKCθ	3	ENST00000397178	16	10	–	6509140	6597286
PKC	ENSG00000163558	*PRKCI*	—	PKCι	2	ENST00000392722	17	3	+	171422906	171506463
PKC	—	—	yes	—	—	—	0	X	–	100679126	100680920
TSC1	ENSG00000165699	*TSC1*	—	Tsc1	2	ENST00000298552	20	9	–	134756557	134809841
FOXO	ENSG00000150907	*FOXO1*	—	FOXO1	1	ENST00000379561	1	13	–	40027801	40138734
FOXO	Pseudogene:2499	—	yes (FOXO1b)	—	—	—	0	5	–	180458749	180460263
FOXO	ENSG00000118689	*FOXO3*	—	FOXO3	1	ENST00000343882	1	6	+	108987719	109112664
FOXO	ENSG00000213688	*AC026271.6*	yes (FOXO3b)	—	1	ENST00000395675	1	17	–	18514942	18515487
FOXO	ENSG00000184481	*FOXO4*	—	FOXO4	4	ENST00000374256	2	X	+	70232751	70238345
FOXO	ENSG00000204060	*FOXO6*	—	FOXO6	2	ENST00000401063	1	1	+	41600190	41620940
GSK3	ENSG00000105723	*GSK3A*	—	GSK3α	2	ENST00000222330	10	19	–	47426186	47438591
GSK3	ENSG00000082701	*GSK3B*	—	GSK3β	2	ENST00000316626	11	3	–	121028238	121295954
TSC2	ENSG00000103197	*TSC2*	—	Tsc2	5	ENST00000219476	40	16	+	2037991	2078713
EIF2BE	ENSG00000145191	*EIF2B5*	—	eIF2Bε	1	ENST00000273783	15	3	+	185335504	185345793
GYS	ENSG00000104812	*GYS1*	—	GYS1	2	ENST00000323798	15	19	–	54163195	54188361
GYS	ENSG00000111713	*GYS2*	—	GYS2	1	ENST00000261195	15	12	–	21580390	21649048
MYC	ENSG00000136997	*MYC*	—	c-Myc	2	ENST00000377970	2	8	+	128817498	128822853
MYC	ENSG00000116990	*MYCL1*	—	L-Myc	3	ENST00000334282	2	1	–	40133685	40140272
MYC	ENSG00000204053	*MYCL2*	—	—	1	ENST00000372451	0	X	+	106402468	106403544
MYC	ENSG00000134323	*MYCN*	—	N-Myc	1	ENST00000281043	1	2	+	15998134	16004579
RHEB	ENSG00000106615	*RHEB*	—	Rheb	1	ENST00000262187	7	7	–	150794032	150847943
RHEB	Pseudogene:4218	—	yes	—	—	—	0	10	+	46334157	46334708
RHEB	Pseudogene:74530	—	yes	—	—	—	0	10	+	46816229	46816781
RHEB	ENSG00000167550	*RHEBL1*	—	Rheb2	1	ENST00000301068	7	12	–	47744735	47750042
TOR	ENSG00000198793	*MTOR*	—	mTOR	2	ENST00000361445	56	1	–	11089180	11245176
4EBP	ENSG00000187840	*EIF4EBP1*	—	4E-BP1	1	ENST00000338825	2	8	+	38007177	38037036
4EBP	Pseudogene:5571	—	yes	—	—	—	0	14	+	20957619	20957956
4EBP	Pseudogene:7248	—	yes	—	—	—	0	22	+	22678527	22678799
4EBP	ENSG00000148730	*EIF4EBP2*	—	4E-BP2	1	ENST00000373218	2	10	+	71833928	71853675
4EBP	—	—	yes	—	—	—	0	6	–	98734099	98734436
4EBP	—	—	yes	—	—	—	0	15	–	40660618	40660978
4EBP	—	—	yes	—	—	—	0	20	–	41485253	41485598
4EBP	ENSG00000131503	*EIF4EBP3*	—	4E-BP3	8	ENST00000310331	2	5	+	139907435	139909347
S6K	ENSG00000108443	*RPS6KB1*	—	S6K1	3	ENST00000225577	14	17	+	55325225	55382564
S6K	Pseudogene:26001	—	yes	—	—	—	10	17	+	20560388	20581029
S6K	Pseudogene:26336	—	yes	—	—	—	8	17	–	57663613	57679296
S6K	ENSG00000175634	*RPS6KB2*	—	S6K2	1	ENST00000312629	14	11	+	66952511	66959454
EIF4E	ENSG00000151247	*EIF4E*	—	eIF4E-1	2	ENST00000394918	6	4	–	100020236	100043647
EIF4E	Pseudogene:1683	—	yes	—	—	—	0	3	–	183660902	183661536
EIF4E	—	—	yes	—	—	—	0	8	+	14210915	14211551
EIF4E	—	—	—	—	—	—	0	17	–	44856633	44857265
EIF4E	Pseudogene:6958	—	yes	—	—	—	0	20	+	5477498	5478124
EIF4E	ENSG00000175766	*EIF4E1B*	—	eIF4E-1b	1	ENST00000318682	4	5	+	176001486	176006248
EIF4E	ENSG00000135930	*EIF4E2*	—	eIF4E-2A	2	ENST00000258416	6	2	+	233123601	233142163
EIF4E	ENSG00000213384	*AC046136.12*	yes	—	1	ENST00000393781	0	3	+	116480662	116481397
EIF4E	ENSG00000163412	*EIF4E3*	—	eIF4E-3	2	ENST00000389826	5	3	–	71814570	71885465

RPS6	ENSG00000137154	*RPS6*	—	RpS6	4	ENST00000380394	5	9	−	19366254	19370235
RPS6	ENSG00000214908	*AL353678.11*	yes	—	1	ENST00000399221	0	9	−	88179178	88180441
RPS6	Pseudogene:918	—	yes	—	—	—	0	2	−	101492454	101493196
RPS6	Pseudogene:1121	—	yes	—	—	—	0	2	+	179575089	179575738
RPS6	Pseudogene:17462	—	yes	—	—	—	0	3	+	164315911	164316665
RPS6	Pseudogene:1826	—	yes	—	—	—	0	4	−	65315546	65316210
RPS6	Pseudogene:18462	—	yes	—	—	—	0	4	−	83635020	83636070
RPS6	—	—	yes	—	—	—	0	6	−	74157485	74158211
RPS6	Pseudogene:2752	—	yes	—	—	—	0	6	+	78262531	78263101
RPS6	Pseudogene:3734	—	yes	—	—	—	0	8	+	104849685	104850430
RPS6	Pseudogene:3803	—	yes	—	—	—	0	9	+	4623029	4623767
RPS6	Pseudogene:3872	—	yes	—	—	—	0	9	−	19190349	19191044
RPS6	Pseudogene:3994	—	yes	—	—	—	0	9	−	84548864	84549530
RPS6	Pseudogene:4223	—	yes	—	—	—	0	10	+	49170748	49171288
RPS6	Pseudogene:4402	—	yes	—	—	—	0	10	−	113247937	113248689
RPS6	—	—	yes	—	—	—	0	11	+	101671001	101671744
RPS6	—	—	yes	—	—	—	0	11	−	111611307	111611971
RPS6	Pseudogene:4895	—	yes	—	—	—	0	12	+	2639364	2639958
RPS6	—	—	yes	—	—	—	0	12	+	12895183	12895938
RPS6	Pseudogene:5113	—	yes	—	—	—	0	12	+	57700217	57700939
RPS6	Pseudogene:24109	—	yes	—	—	—	0	12	−	100146059	100146785
RPS6	Pseudogene:5255	—	yes	—	—	—	0	12	−	130343083	130343674
RPS6	Pseudogene:5254	—	yes	—	—	—	0	12	+	130717990	130718744
RPS6	Pseudogene:24300	—	yes	—	—	—	0	13	+	98623843	98624747
RPS6	Pseudogene:5599	—	yes	—	—	—	0	14	+	29594157	29594908
RPS6	Pseudogene:6748	—	—	—	—	—	0	19	+	12865947	12866691
RPS6	Pseudogene:7508	—	yes	—	—	—	0	X	−	73512688	73513445
CYH	ENSG00000108669	*CYTH1*	—	Cyh1	4	ENST00000361101	12	17	−	74181725	74289971
CYH	ENSG00000105443	*CYTH2*	—	Cyh2	2	ENST00000325139	11	19	+	53664424	53674457
CYH	ENSG00000008256	*CYTH3*	—	Cyh2	2	ENST00000396741	12	7	−	6167939	6278800
CYH	ENSG00000100055	*CYTH4*	—	Cyh3	4	ENST00000248901	12	22	+	36008370	36041326
PTEN	ENSG00000171862	*PTEN*	—	PTEN	1	ENST00000371953	8	10	+	89613175	89718511
PTEN	—	—	—	—	—	—	0	9	−	33665441	33666649

Chr, chromosome.

[a]When available, either the Ensembl or the Pseudogene.org (build 36; Karro et al. 2007) is provided.

[b]HUGO Gene Nomenclature Committee (HGNC; Eyre et al. 2006) symbol.

[c]The pseudogene name is provided if it is described in the literature.

[d]Number of encoded proteins.

[e]Number of introns within the CDS

Table S2. Human organs and tissues used in gene expression analyses.

Tissue
Whole blood
Thymus
Tonsil
Lymph node
Bone marrow
Whole brain
Lung
Heart
Liver
Kidney
Prostate
Uterus
Thyroid
Placenta
Smooth muscle
Adipocyte
Pancreas
Testis
Salivary gland
Trachea
Ovary
Appendix
Skin
Skeletal muscle
Tongue

Tissue/organ names from Su et al. (2004).

Table S3. Dataset 1.

Gene	Pathway position	6 species					5 species[a]					Gene expression		Connectivity	Protein length[f]
		d_N[b]	d_S[b]	ω	ENC	%used codons[c]	d_N[b]	d_S[b]	ω	ENC	%used codons[c]	Level[d]	Breadth[e]		
INSR	0	0.185	4.344	0.043	49.77	77.50	0.155	4.131	0.038	49.52	91.61	1012.74	22	73	1382
IGF1R	0	0.135	3.155	0.043	51.61	78.49	0.114	2.943	0.039	48.43	96.20	869.08	25	50	1367
INSRR	—[g]	0.254	5.072	0.050	42.13	11.80	0.209	3.597	0.058	42.39	57.90	149.04	6	2	1297
IRS1	1	0.068	4.688	0.015	41.98	23.19	0.081	3.096	0.026	42.95	48.07	340.74	18	66	1242
IRS2[h]	1	—	—	—	—	0.00	—	—	—	—	0.00	1435.34	25	39	1338
AC069281.1[h]	1	—	—	—	—	0.00	—	—	—	—	0.00	—	—	0	269
IRS4	1	0.267	11.123	0.024	54.16	10.50	0.206	6.688	0.031	54.56	15.91	303.56	17	13	1257
DOK1[h]	1	—	—	—	—	0.00	—	—	—	—	0.00	480.90	24	33	481
DOK2[h]	1	—	—	—	—	0.00	—	—	—	—	0.00	437.26	12	15	412
DOK3	1	0.404	6.438	0.063	37.66	32.46	0.357	5.043	0.071	38.46	39.72	281.12	17	6	496
DOK4	1	0.116	4.467	0.026	40.49	81.90	0.093	4.020	0.023	41.46	81.90	223.30	13	8	326
DOK5	1	0.088	2.333	0.038	54.14	92.81	0.076	2.011	0.038	54.52	100.00	563.04	25	7	306
DOK6	1	0.024	2.274	0.011	54.75	72.81	0.008	1.675	0.005	54.60	79.46	—	—	1	331
FRS2	1	0.074	2.528	0.029	54.18	77.95	0.084	1.496	0.056	54.40	100.00	68.34	1	20	508
FRS3	1	0.254	4.850	0.052	42.62	56.30	0.171	3.491	0.049	43.48	60.57	542.06	24	10	492
PIK3R1	2	0.131	4.610	0.029	55.14	77.73	0.103	1.912	0.054	55.64	98.63	835.66	25	132	732
PIK3R2	2	0.258	6.953	0.037	40.69	88.05	0.221	6.752	0.033	39.66	89.42	689.32	25	29	728
PIK3R3	2	0.116	3.818	0.030	56.02	79.18	0.103	3.262	0.032	56.05	79.18	566.02	25	12	461
PIK3CA	3	0.032	2.163	0.015	52.15	99.72	0.030	1.820	0.016	51.34	99.72	143.28	7	19	1068
PIK3CB	3	0.129	2.430	0.053	52.82	95.05	0.107	2.146	0.050	52.71	99.25	419.30	24	7	1070
PIK3CG	—[g]	0.227	4.765	0.048	52.62	86.03	0.148	2.683	0.055	52.73	98.19	82.16	2	16	1102
PIK3CD	3	0.157	5.097	0.031	37.40	86.58	0.129	4.641	0.028	37.24	89.26	1253.48	25	12	1044
VEPH1	4	0.338	2.774	0.122	53.15	92.80	0.278	2.235	0.124	52.62	92.80	80.54	2	4	833
PDPK1	4	0.025	2.916	0.009	52.72	44.06	0.077	2.064	0.038	52.54	80.04	1338.84	25	36	556
AKT1	5	0.051	6.026	0.009	45.77	74.38	0.042	4.984	0.008	40.99	92.71	970.02	19	108	480
AKT2	—[g]	0.114	6.846	0.017	43.01	58.21	0.099	6.496	0.015	45.38	93.76	283.36	21	19	481
AKT3	5	0.006	1.457	0.004	53.60	94.15	0.005	1.127	0.004	53.36	100.00	1041.90	25	5	479
PRKCA	5	0.040	2.899	0.014	53.17	97.62	0.035	2.481	0.014	52.82	97.62	557.32	25	168	672
PRKCB	5	0.029	2.663	0.011	49.98	80.83	0.036	2.354	0.015	49.26	89.90	1064.62	24	66	673
PRKCG[h]	5	—	—	—	—	0.00	—	—	—	—	0.00	35.90	0	48	697
PRKCD	5	0.152	3.832	0.040	48.04	68.79	0.141	3.278	0.043	45.15	95.56	413.02	16	97	676
PRKCE	5	0.037	4.613	0.008	51.22	58.34	0.035	2.263	0.015	52.33	98.64	685.74	25	43	737
PRKCZ	5	0.095	2.856	0.033	52.45	95.78	0.084	2.387	0.035	49.65	95.95	1106.96	15	68	592
PRKCH	5	0.157	4.082	0.038	51.62	77.16	0.109	2.788	0.039	51.64	99.71	867.08	25	6	683
PRKCQ	5	0.1467	3.372	0.0435	54.21	86.97	0.932	22.953	0.041	54.33	86.97	1104.80	25	23	706
PRKCI	5	0.017	2.722	0.006	55.98	88.09	0.029	2.208	0.013	56.39	95.13	1147.88	25	33	596
TSC1	5	0.275	1.970	0.140	54.69	91.49	0.228	1.708	0.134	54.54	91.49	715.78	25	15	1164
FOXO1	6	0.202	3.688	0.055	47.50	74.35	0.157	2.419	0.065	49.32	87.63	868.36	25	23	655
FOXO3	6	0.134	3.413	0.039	42.31	72.07	0.095	2.891	0.033	43.15	76.23	1289.90	25	15	673
FOXO4	6	0.053	4.673	0.011	50.71	7.33	0.163	3.277	0.050	50.84	47.33	233.86	10	10	505
FOXO6[h]	6	—	—	—	—	0.00	—	—	—	—	0.00	—	—	0	492
GSK3A[h]	6	—	—	—	—	0.00	—	—	—	—	0.00	1617.88	25	21	483
GSK3B	6	0.003	1.551	0.002	53.88	62.59	0.027	0.974	0.027	53.99	99.77	561.56	25	88	433
TSC2	6	0.179	4.578	0.039	45.64	83.45	0.144	3.032	0.048	49.98	94.08	358.20	17	22	1807
EIF2B5	7	0.285	3.861	0.074	53.98	86.82	0.235	2.876	0.082	54.04	91.40	711.14	25	68	721
GYS1	7	0.078	5.856	0.013	41.29	50.07	0.069	7.092	0.010	42.24	50.07	1352.30	25	10	737
GYS2	7	0.158	3.991	0.040	52.53	47.80	0.140	2.291	0.061	52.57	97.30	131.14	3	1	703
MYC	7	0.198	3.513	0.057	43.58	47.58	0.150	3.056	0.049	44.20	73.57	490.34	18	148	454
MYCL1	7	0.182	4.352	0.042	46.38	37.31	0.170	4.442	0.038	48.12	37.31	550.40	25	2	394
MYCN[h]	7	—	—	—	—	0.00	—	—	—	—	0.00	94.26	1	7	464
RHEB	7	0.017	1.481	0.011	45.33	90.22	0.018	1.115	0.016	46.09	100.00	2303.66	25	7	184
RHEBL1[h]	7	—	—	—	—	0.00	—	—	—	—	0.00	191.68	10	7	183
MTOR	8	0.027	2.916	0.009	49.47	94.00	0.023	2.374	0.010	50.68	96.43	302.60	20	15	2549
EIF4EBP1	9	0.171	3.370	0.051	43.46	38.14	0.116	2.849	0.041	45.02	68.64	558.70	19	15	118
EIF4EBP2	9	0.078	4.217	0.019	52.27	30.83	0.115	2.316	0.050	53.60	75.83	679.22	25	1	120
EIF4EBP3[h]	9	—	—	—	—	0.00	—	—	—	—	0.00	—	—	4	100
RPS6KB1	9	0.019	1.188	0.016	53.83	89.90	0.014	1.020	0.013	53.44	94.48	164.44	7	21	525
RPS6KB2	9	0.103	4.120	0.025	41.72	33.82	0.084	3.225	0.026	43.49	33.82	350.30	19	6	482
EIF4E	10	0.036	1.081	0.033	54.81	94.93	0.031	0.815	0.038	54.20	94.93	—	—	106	217
EIF4E1B	10	0.338	4.440	0.076	42.38	96.00	0.288	3.672	0.079	39.94	99.43	—	—	0	217
EIF4E2	—[g]	0.075	2.080	0.036	55.64	48.16	0.031	2.063	0.015	56.71	90.20	639.70	25	6	245
EIF4E3	—[g]	0.141	2.617	0.054	57.33	100.00	0.106	2.575	0.041	57.57	100.00	423.74	24	0	165
RPS6	10	0.011	4.795	0.002	50.80	81.53	0.011	2.570	0.004	51.18	99.20	23490.48	25	179	249
CYTH1	—[g]	0.045	4.394	0.010	48.99	79.40	0.028	3.248	0.009	49.98	96.98	466.28	17	8	398
CYTH2[h]	—[g]	—	—	—	—	0.00	—	—	—	—	0.00	599.10	24	15	400
CYTH3	—[g]	0.029	3.371	0.009	51.76	87.25	0.026	3.327	0.008	51.06	87.25	38.94	1	5	400
CYTH4	—[g]	0.226	6.380	0.035	40.66	97.97	0.188	5.532	0.034	42.97	97.97	693.06	25	0	394
PTEN	—[g]	0.037	1.392	0.026	52.57	93.05	0.032	1.183	0.027	52.64	93.05	797.40	24	31	403

ENC, effective number of codons (Wright 1990).
[a]Excluding *O. anatinus* from the analyses.
[b]Values estimated as the sum across all branches of the phylogeny.
[c]Percent of used codons for estimating d_N, d_S and ω values.
[d]Expression levels averaged across 25 selected human tissues/organs (table S2).
[e]Number of selected tissues with expression level ≥ 200.
[f]Number of amino acids of the human protein.
[g]Genes excluded from network-level analyses.
[h]Genes not present in all six studied species.

Table S4. Dataset 3.

Paralogous group	Pathway position	6 species					5 species[a]					Gene expression		Connectivity	Protein length[f]	#copies[g]
		d_N[b]	d_S[b]	ω	ENC	%used codons[c]	d_N[b]	d_S[b]	ω	ENC	%used codons[c]	Level[d]	Breadth[e]			
INR	0	0.160	3.750	0.043	50.69	77.99	0.134	3.537	0.038	48.97	93.90	940.91	23.5	61.5	1374.5	3
IRS	1	0.162	4.838	0.032	47.50	37.33	0.134	3.440	0.037	48.05	43.80	467.57	17.6	18.2	621.5	11
P85	2	0.169	5.127	0.032	50.62	81.65	0.142	3.975	0.039	50.45	89.08	697.00	25.0	57.7	640.3	3
P110	3	0.106	3.230	0.033	47.46	93.78	0.089	2.869	0.031	47.10	96.08	605.35	18.7	12.7	1060.7	4
MELT	4	0.338	2.774	0.122	53.15	92.80	0.278	2.235	0.124	52.62	92.80	80.54	2.0	4.0	833.0	1
PDK1	4	0.025	2.916	0.009	52.72	44.06	0.077	2.064	0.038	52.54	80.04	1338.84	25.0	36.0	556.0	2
AKT	5	0.028	3.742	0.006	49.69	84.26	0.023	3.055	0.006	47.17	96.35	1005.96	22.0	56.5	479.5	3
PKC	5	0.084	3.380	0.024	52.08	72.62	0.175	5.089	0.027	51.45	84.39	775.92	20.0	61.3	670.2	9
TSC1	5	0.275	1.970	0.140	54.69	91.49	0.228	1.708	0.134	54.54	91.49	715.78	25.0	15.0	1164.0	1
FOXO	6	0.130	3.925	0.035	46.84	38.44	0.138	2.862	0.049	47.77	52.80	797.37	20.0	12.0	581.3	4
GSK3	6	0.003	1.551	0.002	53.88	31.29	0.027	0.974	0.027	53.99	49.88	1089.72	25.0	54.5	458.0	2
TSC2	6	0.179	4.578	0.039	45.64	83.45	0.144	3.032	0.048	49.98	94.08	358.20	17.0	22.0	1807.0	1
EIF2Bε	7	0.285	3.861	0.074	53.98	86.82	0.235	2.876	0.082	54.04	91.40	711.14	25.0	68.0	721.0	1
GYS	7	0.118	4.923	0.026	46.91	48.93	0.104	4.692	0.035	47.40	73.68	741.72	14.0	5.5	720.0	2
MYC	7	0.190	3.932	0.049	44.98	28.30	0.160	3.749	0.044	46.16	36.96	378.33	14.7	52.3	437.3	4
RHEB	7	0.017	1.481	0.011	45.33	45.11	0.018	1.115	0.016	46.09	50.00	1247.67	17.5	7.0	183.5	2
TOR	8	0.027	2.916	0.009	49.47	94.00	0.023	2.374	0.010	50.68	96.43	302.60	20.0	15.0	2549.0	1
4EBP	9	0.124	3.793	0.035	47.86	22.99	0.115	2.582	0.045	49.31	48.16	618.96	22.0	6.7	112.7	3
S6K	9	0.061	2.654	0.021	47.78	61.86	0.049	2.123	0.020	48.47	64.15	257.37	13.0	13.5	503.5	2
EIF4E	10	0.187	2.760	0.055	48.59	95.47	0.160	2.243	0.058	47.07	97.18	—	—	53.0	217.0	5
RPS6	10	0.011	4.795	0.002	50.80	81.53	0.011	2.570	0.004	51.18	99.20	23490.48	25.0	179.0	249.0	2
CYH	—[h]	0.100	4.715	0.018	47.14	66.15	0.081	4.036	0.017	48.00	70.55	449.35	16.8	7.0	398.0	4
PTEN	—[h]	0.037	1.392	0.026	52.57	93.05	0.032	1.183	0.027	52.64	93.05	797.40	24.0	31.0	403.0	2

ENC, effective number of codons (Wright 1990).
[a]Excluding *O. anatinus* from the analyses.
[b]Values estimated as the sum across all branches of the phylogeny.
[c]Percent of used codons for estimating d_N, d_S and ω values.
[d]Expression levels averaged across 25 selected human tissues/organs (table S2).
[e]Number of selected tissues with expression level ≥ 200.
[f]Number of amino acids of the human protein.
[g]Number of putatively functional copies within the paralogous group.
[h]Paralogous groups excluded from network-level analyses.

Table S5. Copy number of the insulin/TOR signaling pathway genes in vertebrates.

	H. sapiens			*M. musculus*			*B. taurus*			*M. domestica*			*O. anatinus*			*G. gallus*		
Symbol	**PF**	**PSE**	**IL**	**PF**	**PSE**	**IL**	**PF**	**PSE**	**IL**	**PF**	**PSE**	**IL**	**PF**	**PSE**	**IL**	**PF**	**PSE**	**IL**
INSR	1			1			1			1			1			1		
IGF1R	1			1			1			1			1			2		
INSRR	1			1			1			1			1			1		
IRS1	1			1			1			1			1			1		
IRS2	1			1			1			1						1		
AC069281.1		1		1			1			1			1					
IRS4	1			1			1			1			1			1		
DOK1	1	1		1			1			1						1		
DOK2	1			1			1			1			1					
DOK3	1			1			1			1			1			1		
DOK4	1			1			1			1			1			1		
DOK5	1			1			1			1			1			1		
DOK6	1			1			1	1		1			1			1		
FRS2	1			1			1			1			1			1		
FRS3	1			1			1			1			1			1		
PIK3R1	1			1			1			1			1			1		
PIK3R2	1			1			1			1			1			2		
PIK3R3	1			1			1			1			1			1		
PIK3CA	1			1			1			1	1		2			2		
PIK3CB	1			1	1		1			1	1	1	1			1		
PIK3CG	1			1			1			1			3			1		
PIK3CD	1			1			1			1			1			1		
VEPH1	1			1			1			1			1			1		
PDPK1	2			1			1			1	1		1			1		
AKT1	1			1			1			1			1			1[a]		
AKT2	1			1			1		1	1			1			1[a]		
AKT3	1			1			1			1			1			1		
PRKCA	1			1			1			1			1			1		
PRKCB	1			1			1			1			1			1		
PRKCG	1			1			1			1			1					
PRKCD	1			1			1			1			1			1		
PRKCE	1			1			1			1			1			1		
PRKCZ	1			1	1		1			1	10	4	1		2	1		
PRKCH	1			1			1			1	2		1			1		
PRKCQ	1			1			1			1			2			2		
PRKCI	1	1		1		1	1			1			1			1		
TSC1	1			1			1			1	1		1			1		
FOXO1	1	1		1			1			1			1			1		
FOXO3	1	1		1			1			1			1			1		
FOXO4	1			1			1			1			1			1		
FOXO6	1			1			1			1			1					
GSK3A	1			1	1		1	1					1					
GSK3B	1			1			1			1			1			1		
TSC2	1			1			1			1			1			1		
EIF2B5	1			1			1			1			1			1		
GYS1	1			1			1			1	1		1			1		
GYS2	1			1			1	1		1			1			1		
MYC	1			1		1	1			1			1			1		
MYCL1	1		1	1			1			1			1			1		
MYCN	1			1		2	1						1			1		
RHEB	1	2		1	1		1		1	1			1			1		
RHEBL1	1			1			1						1					
MTOR	1			1			1			1			1			1		

MTOR	1			1			1			1			1			1	
EIF4EBP1	1	2		1	2		1			1			1			1	
EIF4EBP2	1	3		1	1		1			1			1			1	
EIF4EBP3	1			1				1			1		1				
RPS6KB1	1	2		1	1		1	1		1	2		2			1	
RPS6KB2	1			1			1			1	2		1			1	
EIF4E	1	3	1	1	10		1	1		1	2	2	1			1	
EIF4E1B	1			1		1	1	1		1	3	2	1			1	
EIF4E2	1	1		1	1		1	1	1	1	2	1	1			1	
EIF4E3	1			1			1			1			1			1	
RPS6	1	25	1	1	152	13	1	31	5	1	4	1	1			1	
CYTH1	1			1			1			1			1			1	
CYTH2	1			1			1	1			1		1				
CYTH3	1			1			1			1			1			1	
CYTH4	1			1			1			1			2			1	
PTEN	1		1	1			1			1			1	1		1	

PF, putatively functional genes; PSE, putative pseudogenes; IL, intronless sequences (putative processed copies) without any pseudogenization footprint. Zero values are not shown.
[a]Gene shared by two orthologous groups.

Table S6. Positive selection tests for the insulin/TOR pathway genes.

	M2a vs. M1a test				M8 vs. M7 test				
Gene	ℓ_{M2a}	ℓ_{M1a}	$2\Delta\ell$	*P*	ℓ_{M8}	ℓ_{M7}	$2\Delta\ell$	*P*	
INSR	-11490.90	-11490.90	0.00	1.000	-11426.56	-11427.34	1.55	0.461	
IGF1R	-10506.35	-10506.35	0.00	1.000	-10464.95	-10465.33	0.75	0.687	
INSRR	-1670.33	-1670.33	0.00	1.000	-1664.52	-1664.84	0.64	0.725	
IRS1	-2646.81	-2646.81	0.00	1.000	-2638.55	-2638.55	0.00	1.000	
IRS2 [a]	—	—	—	—	—	—	—	—	
AC069281.1 [a]	—	—	—	—	—	—	—	—	
IRS4	-1369.32	-1369.32	0.00	1.000	-1369.55	-1372.87	6.63	0.036	*
DOK1 [a]	—	—	—	—	—	—	—	—	
DOK2 [a]	—	—	—	—	—	—	—	—	
DOK3	-1957.04	-1957.04	0.00	1.000	-1946.68	-1946.68	0.00	1.000	
DOK4	-2427.05	-2427.05	0.00	1.000	-2415.23	-2415.25	0.03	0.987	
DOK5	-2585.26	-2585.26	0.00	1.000	-2582.88	-2582.88	0.00	1.000	
DOK6	-1995.08	-1995.08	0.00	1.000	-1993.02	-1993.02	0.00	1.000	
FRS2	-3482.07	-3482.07	0.00	1.000	-3478.87	-3478.87	0.00	1.000	
FRS3	-3042.52	-3042.52	0.00	1.000	-3025.43	-3025.43	0.00	1.000	
PIK3R1	-5523.56	-5523.56	0.00	1.000	-5496.20	-5496.20	0.00	1.000	
PIK3R2	-7084.82	-7084.82	0.00	1.000	-7023.50	-7023.50	0.00	1.000	
PIK3R3	-3402.18	-3402.18	0.00	1.000	-3396.87	-3398.16	2.58	0.275	
PIK3CA	-8584.50	-8584.50	0.00	1.000	-8574.35	-8574.35	0.00	1.000	
PIK3CB	-9552.21	-9552.21	0.00	1.000	-9526.22	-9526.22	0.00	1.000	
PIK3CG	-10288.54	-10288.54	0.00	1.000	-10240.46	-10241.38	1.84	0.399	
PIK3CD	-8579.80	-8579.80	0.00	1.000	-8547.97	-8548.55	1.16	0.559	
VEPH1	-8857.68	-8857.68	0.00	1.000	-8828.79	-8828.88	0.18	0.915	
PDPK1	-2102.82	-2102.82	0.00	1.000	-2102.45	-2102.45	0.00	1.000	
AKT1	-3272.95	-3272.95	0.00	1.000	-3264.87	-3266.11	2.49	0.288	
AKT2	-2640.01	-2640.01	0.00	1.000	-2636.19	-2636.25	0.11	0.949	
AKT3	-3120.45	-3120.45	0.00	1.000	-3120.46	-3123.84	6.76	0.034	*
PRKCA	-5716.83	-5716.83	0.00	1.000	-5695.15	-5695.15	0.00	1.000	
PRKCB	-4496.76	-4496.76	0.00	1.000	-4486.57	-4486.57	0.00	1.000	
PRKCG [a]	—	—	—	—	—	—	—	—	
PRKCD	-4685.31	-4685.31	0.00	1.000	-4665.37	-4668.65	6.57	0.038	*
PRKCE	-3661.59	-3661.59	0.00	1.000	-3651.07	-3651.07	0.00	1.000	
PRKCZ	-5316.62	-5316.62	0.00	1.000	-5306.94	-5308.64	3.40	0.183	
PRKCH	-5239.53	-5239.53	0.00	1.000	-5212.12	-5215.11	5.99	0.050	
PRKCQ	-6171.45	-6171.45	0.00	1.000	-6145.33	-6145.33	0.00	1.000	
PRKCI	-4262.39	-4262.39	0.00	1.000	-4259.26	-4259.26	0.00	1.000	
TSC1	-11453.27	-11453.27	0.00	1.000	-11441.52	-11442.27	1.50	0.472	
FOXO1	-5112.86	-5112.86	0.00	1.000	-5081.28	-5081.28	0.00	1.000	
FOXO3	-4530.82	-4530.82	0.00	1.000	-4497.30	-4497.30	0.00	1.000	
FOXO4	-277.62	-277.62	0.00	1.000	-276.91	-276.91	0.00	1.000	
FOXO6 [a]	—	—	—	—	—	—	—	—	
GSK3A [a]	—	—	—	—	—	—	—	—	
GSK3B	-1991.81	-1991.81	0.00	1.000	-1991.81	-1994.24	4.85	0.088	

TSC2	-16179.94	-16179.94	0.00	1.000	-16104.76	-16106.36	3.20	0.202
EIF2B5	-6958.42	-6958.42	0.00	1.000	-6933.60	-6935.16	3.11	0.211
GYS1	-3247.07	-3247.07	0.00	1.000	-3232.24	-3232.24	0.00	1.000
GYS2	-3456.03	-3456.03	0.00	1.000	-3441.96	-3441.96	0.00	1.000
MYC	-2257.73	-2257.73	0.00	1.000	-2251.57	-2251.83	0.53	0.766
MYCL1	-1493.49	-1493.49	0.00	1.000	-1487.29	-1487.30	0.02	0.991
MYCN [a]	—	—	—	—	—	—	—	—
RHEB	-1230.37	-1230.37	0.00	1.000	-1229.68	-1229.68	0.00	1.000
RHEBL1 [a]	—	—	—	—	—	—	—	—
MTOR	-20513.67	-20513.67	0.00	1.000	-20479.01	-20480.39	2.75	0.253
EIF4EBP1	-408.09	-408.09	0.00	1.000	-408.05	-408.09	0.08	0.959
EIF4EBP2	-327.01	-327.01	0.00	1.000	-326.69	-326.69	0.00	1.000
EIF4EBP3 [a]	—	—	—	—	—	—	—	—
RPS6KB1	-3333.45	-3333.45	0.00	1.000	-3332.97	-3334.93	3.91	0.141
RPS6KB2	-1438.28	-1438.28	0.00	1.000	-1428.24	-1428.24	0.00	1.000
EIF4E	-1452.06	-1452.06	0.00	1.000	-1451.63	-1451.72	0.19	0.909
EIF4E1B	-1931.58	-1931.58	0.00	1.000	-1916.89	-1916.89	0.00	1.000
EIF4E2	-959.40	-959.40	0.00	1.000	-954.97	-954.97	0.00	1.000
EIF4E3	-1512.86	-1512.86	0.00	1.000	-1509.62	-1509.62	0.00	1.000
RPS6	-1811.48	-1811.48	0.00	1.000	-1811.65	-1811.65	0.00	1.000
CYTH1	-2764.74	-2764.74	0.00	1.000	-2756.12	-2756.40	0.56	0.755
CYTH2 [a]	—	—	—	—	—	—	—	—
CYTH3	-3043.91	-3043.91	0.00	1.000	-3042.16	-3042.16	0.00	1.000
CYTH4	-4135.86	-4135.86	0.00	1.000	-4113.70	-4116.64	5.87	0.053
PTEN	-2693.61	-2693.61	0.00	1.000	-2691.57	-2691.57	0.00	1.000

ℓ_i, log-likelihood of the observed data under the evolutionary model *i*. *, $P < 0.05$.
[a]Genes not present in all six studied species.

Table S7. Similarity in ω, d_N and d_S values among genes encoding interacting proteins.

		ω		d_N		d_S	
Dataset	n [a]	X	P	X	P	X	P
2	33	0.023	0.002 **	0.079	0.005 **	1.934	0.873
	20[b]	0.028	0.047 *	0.088	0.051	2.199	0.944
3	33	0.024	0.012 *	0.084	0.028 *	1.316	0.702
	20[b]	0.027	0.062	0.088	0.063	1.439	0.746

*, $P < 0.05$; **, $P < 0.01$.
[a]Number of interactions considered.
[b]Analysis considering only the interactions used in *Drosophila* (Alvarez-Ponce et al. 2009).

Table S8. Bivariate correlations (dataset 1).

	Position		ω		d_N		d_S		ENC		%used codons		Expression level		Exp. breadth		Connectivity		Protein length	
	ρ	P	ρ	P	ρ	P	ρ	P	ρ	P	ρ	P	ρ	P	ρ	P	ρ	P	ρ	P
Position	—	—	-0.023	0.438	-0.134	0.182	-0.143	0.334	-0.073	0.620	0.021	0.885	0.110	0.470	0.091	0.554	-0.014	0.923	-0.365	0.011 *
ω	-0.073	0.312	—	—	0.796	<0.001 ***	0.069	0.640	0.050	0.736	-0.094	0.526	-0.315	0.035 *	-0.171	0.262	-0.176	0.230	0.088	0.551
d_N	-0.162	0.136	0.887	<0.001 ***	—	—	0.589	<0.001 ***	-0.219	0.134	-0.430	0.002 **	-0.214	0.158	-0.121	0.428	-0.192	0.190	0.231	0.114
d_S	-0.128	0.387	0.085	0.567	0.447	0.001 **	—	—	-0.594	<0.001 ***	-0.601	<0.001 ***	0.039	0.799	-0.043	0.781	-0.035	0.815	0.172	0.242
ENC	-0.109	0.462	-0.142	0.336	-0.276	0.058	-0.558	<0.001 *	—	—	0.312	0.031 *	-0.067	0.660	0.165	0.279	0.053	0.719	0.019	0.901
%used codons	0.018	0.904	0.050	0.736	-0.160	0.277	-0.533	<0.001 *	0.305	0.035 *	—	—	0.092	0.547	0.101	0.509	0.148	0.316	0.025	0.868
Expression level	0.110	0.470	-0.220	0.147	-0.219	0.147	0.000	0.998	-0.012	0.939	0.138	0.364	—	—	0.745	<0.001 ***	0.288	0.055	-0.080	0.601
Expression breadth	0.091	0.554	-0.137	0.369	-0.125	0.414	-0.015	0.922	0.144	0.346	0.101	0.511	0.745	<0.001 ***	—	—	0.097	0.526	-0.099	0.518
Connectivity	-0.014	0.923	-0.189	0.199	-0.206	0.161	-0.009	0.951	0.139	0.344	0.096	0.514	0.288	0.055	0.097	0.526	—	—	0.200	0.173
Protein length	-0.365	0.011 *	0.153	0.298	0.265	0.069	0.173	0.240	0.011	0.942	0.106	0.471	-0.080	0.601	-0.099	0.518	0.200	0.173	—	—

*, $P < 0.05$; **, $P < 0.01$; ***, $P < 0.001$. Values below and above the principal diagonal are based on analyses using all six species or excluding *O. anatinus* from the analyses, respectively. All correlations are based on $n = 48$ observations except those involving gene expression level and breadth ($n = 45$). Two-tailed *P*-values are provided except for the correlations between pathway position and ω and d_N (one-tailed).

Table S9. Bivariate correlations (dataset 3).

	Position		ω		d_N		d_S		ENC		%used codons		Expression level		Exp. breadth		Connectivity		Protein length	
	ρ	P	ρ	P	ρ	P	ρ	P	ρ	P	ρ	P	ρ	P	ρ	P	ρ	P	ρ	P
Position	—	—	-0.154	0.252	-0.245	0.142	-0.293	0.198	-0.152	0.512	0.059	0.800	-0.030	0.902	-0.127	0.593	-0.081	0.727	-0.483	0.026 *
ω	-0.134	0.281	—	—	0.848	<0.001 ***	-0.016	0.947	0.240	0.294	-0.099	0.670	-0.337	0.146	0.016	0.947	-0.173	0.454	0.227	0.322
d_N	-0.198	0.195	0.943	<0.001 ***	—	—	0.294	0.197	0.284	0.211	-0.058	0.801	-0.439	0.053	-0.097	0.684	0.032	0.891	0.360	0.109
d_S	-0.137	0.555	0.071	0.758	0.247	0.281	—	—	-0.251	0.273	-0.023	0.920	-0.144	0.544	-0.164	0.489	0.286	0.208	0.252	0.271
ENC	-0.209	0.364	0.009	0.969	0.042	0.858	-0.287	0.207	—	—	0.138	0.552	0.042	0.860	0.547	0.012 *	0.271	0.235	0.375	0.094
%used codons	-0.022	0.924	0.281	0.218	0.283	0.214	-0.214	0.351	0.303	0.182	—	—	0.027	0.910	0.215	0.362	0.323	0.153	0.383	0.086
Expression level	-0.030	0.902	-0.498	0.026 *	-0.570	0.009 **	-0.062	0.796	0.236	0.316	-0.274	0.243	—	—	0.606	0.005 **	0.393	0.087	-0.417	0.068
Expression breadth	-0.127	0.593	-0.208	0.378	-0.195	0.409	0.008	0.975	0.650	0.002 **	0.048	0.840	0.606	0.005 **	—	—	0.631	0.003 **	-0.079	0.742
Connectivity	-0.081	0.727	-0.168	0.468	-0.069	0.767	0.221	0.336	0.433	0.050	0.121	0.600	0.393	0.087	0.631	0.003 **	—	—	-0.021	0.927
Protein length	-0.483	0.026 *	0.334	0.139	0.345	0.125	0.108	0.642	0.205	0.372	0.522	0.015 *	-0.417	0.068	-0.079	0.742	-0.021	0.927	—	—

*, $P < 0.05$; **, $P < 0.01$; ***, $P < 0.001$. Values below and above the principal diagonal are based on analyses using all six species or excluding *O. anatinus* from the analyses, respectively. All correlations are based on $n = 21$ observations except those involving gene expression level and breadth (for which $n = 20$). Two-tailed P-values are provided except for the correlations between pathway position and ω and d_N.

Table S10. Correlations between pathway position and ω, d_N and d_S for each phylogenetic branch (dataset 1).

			ω		d_N		d_S	
# species	Branch[a]	*n*	ρ	*P* [b]	ρ	*P* [b]	ρ	*P* [c]
6	all[d]	48	-0.073	0.312	-0.162	0.136	-0.128	0.387
	a	48	0.139	0.828	0.122	0.797	-0.001	0.993
	b	48	-0.120	0.208	-0.191	0.097	-0.293	0.043 *
	c	47	-0.103	0.246	-0.169	0.128	-0.296	0.043 *
	d	47	0.003	0.508	-0.163	0.137	-0.138	0.354
	e	47	-0.177	0.117	-0.209	0.079	-0.087	0.562
	f	46	-0.100	0.255	-0.063	0.339	-0.053	0.728
	g	28	-0.513	0.003 **	-0.444	0.009 **	0.215	0.272
	h	48	-0.329	0.011 *	-0.272	0.031 *	-0.096	0.515
	i	35	-0.223	0.099	-0.231	0.092	0.202	0.245
5[e]	all[d]	48	-0.023	0.438	-0.134	0.182	-0.143	0.334
	a	48	-0.070	0.318	-0.063	0.335	-0.054	0.714
	b	48	0.004	0.512	-0.128	0.192	-0.283	0.052
	c	48	-0.120	0.208	-0.207	0.079	-0.287	0.048 *
	d	46	-0.048	0.377	-0.209	0.081	-0.140	0.353
	f+i	47	-0.043	0.387	-0.027	0.429	-0.003	0.984
	g	26	-0.367	0.033 *	-0.410	0.019 *	-0.017	0.936
	h	47	-0.318	0.015 *	-0.244	0.049 *	-0.082	0.583

*, $P < 0.05$; **, $P < 0.01$.
[a]Branch codes according to figure 1.
[b]One-tailed *P*-values.
[c]Two-tailed *P*-values.
[d]Using overall ω, d_N and d_S values.
[e]Excluding *O. anatinus* from the analysis.

Table S11. Correlations between pathway position and ω, d_N and d_S for each phylogenetic branch (dataset 3).

			ω		d_N		d_S	
# species	Branch[a]	*n*	ρ	P[b]	ρ	P[b]	ρ	P[c]
6	all[d]	21	-0.134	0.281	-0.198	0.195	-0.137	0.555
	a	21	-0.001	0.499	-0.084	0.359	0.008	0.973
	b	21	-0.178	0.221	-0.217	0.173	-0.346	0.125
	c	21	-0.115	0.310	-0.195	0.198	-0.466	0.033 *
	d	21	-0.225	0.163	-0.226	0.163	-0.127	0.582
	e	21	-0.250	0.137	-0.237	0.150	-0.020	0.931
	f	21	-0.271	0.118	-0.195	0.199	-0.014	0.953
	g	18	-0.498	0.018 *	-0.510	0.015 *	0.238	0.342
	h	21	-0.516	0.008 *	-0.548	0.005 *	-0.199	0.386
	i	18	-0.283	0.128	-0.269	0.140	0.302	0.223
5[e]	all[d]	21	-0.154	0.252	-0.245	0.142	-0.293	0.198
	a	21	-0.290	0.101	-0.314	0.083	-0.122	0.598
	b	21	-0.201	0.191	-0.245	0.142	-0.329	0.145
	c	21	-0.218	0.171	-0.326	0.075	-0.507	0.019 *
	d	21	-0.362	0.054	-0.250	0.137	-0.189	0.412
	f+i	21	-0.137	0.276	-0.105	0.325	0.007	0.975
	g	18	-0.530	0.012 *	-0.587	0.005 *	-0.010	0.967
	h	21	-0.390	0.040 *	-0.510	0.009 *	-0.246	0.281

*, $P < 0.05$.
[a]Branch codes according to figure 1.
[b]One-tailed *P*-values.
[c]Two-tailed *P*-values.
[d]Using overall ω, d_N and d_S values.
[e]Excluding *O. anatinus* from the analysis.

Table S12. Path analysis.

			ω		d_N		d_S	
#species	Dataset	*n*	*β*	*P* [a]	*β*	*P* [a]	*β*	*P* [b]
6	1	45	-0.132	0.194	-0.246	0.041 *	-0.227	0.109
	2	20	-0.138	0.276	-0.189	0.213	-0.156	0.380
	3	20	-0.125	0.267	-0.198	0.167	-0.147	0.481
5[c]	1	45	-0.145	0.170	-0.196	0.104	-0.126	0.406
	2	20	-0.307	0.074	-0.307	0.090	-0.033	0.847
	3	20	-0.298	0.033 *	-0.322	0.042 *	-0.094	0.630

Association between pathway position and ω, d_N and d_S values after controlling for expression level, codon bias, protein length and connectivity. The analysis conducted using expression breadth instead of expression level yielded equivalent results. *, $P < 0.05$.
[a]One-tailed *P*-values.
[b]Two-tailed *P*-values.
[c]Excluding *O. anatinus* sequences from the analysis.

Table S13. Partial correlation analysis.

			ω		d_N		d_S	
#species	Dataset	*n*	ρ	P [a]	ρ	P [a]	ρ	P [b]
6	1	45	-0.127	0.215	-0.176	0.135	-0.175	0.272
	2	20	-0.144	0.299	-0.117	0.335	0.084	0.762
	3	20	-0.214	0.215	-0.314	0.117	-0.089	0.747
5[c]	1	45	-0.106	0.256	-0.149	0.176	-0.167	0.296
	2	20	-0.408	0.054	-0.321	0.111	0.236	0.381
	3	20	-0.310	0.120	-0.405	0.055	-0.194	0.477

Association between pathway position and ω, d_N and d_S values after controlling for expression level and breadth, codon bias, protein length and connectivity.
[a]One-tailed *P*-values.
[b]Two-tailed *P*-values.
[c]Excluding *O. anatinus* sequences from the analysis.

Table S14. Partial correlation and path analysis (dataset 1)

			Partial correlation analysis						Path analysis					
			ω		d_N		d_S		ω		d_N		d_S	
# species	Branch[a]	*n*	ρ	P[b]	ρ	P[b]	ρ	P[c]	β	P[b]	β	P[b]	β	P[c]
6	all[d]	45	-0.127	0.215	-0.176	0.135	-0.175	0.272	-0.132	0.194	-0.246	0.041 *	-0.227	0.109
	a	45	0.122	0.776	0.144	0.815	0.080	0.621	0.142	0.813	0.146	0.820	0.069	0.587
	b	45	-0.174	0.139	-0.207	0.096	-0.448	0.002 **	-0.093	0.276	-0.233	0.052	-0.313	0.010 *
	c	44	-0.081	0.311	-0.149	0.179	-0.245	0.123	-0.102	0.258	-0.145	0.185	-0.198	0.157
	d	44	-0.055	0.369	-0.154	0.171	-0.173	0.287	-0.038	0.404	-0.172	0.128	-0.233	0.119
	e	44	-0.128	0.216	-0.115	0.241	-0.042	0.800	-0.133	0.192	-0.167	0.117	0.039	0.808
	f	43	-0.124	0.227	-0.030	0.428	0.006	0.969	-0.153	0.156	-0.114	0.211	0.063	0.654
	g	27	-0.384	0.032 *	-0.357	0.044 *	0.163	0.460	-0.525	0.002 **	-0.488	0.001 **	0.007	0.974
	h	45	-0.319	0.019 *	-0.258	0.050 *	-0.128	0.426	-0.258	0.027 *	-0.341	0.007 **	-0.138	0.356
	i	32	-0.278	0.074	-0.353	0.030 *	-0.028	0.889	-0.212	0.043 *	-0.350	0.006 **	0.045	0.764
5[e]	all[d]	45	-0.106	0.256	-0.149	0.176	-0.167	0.296	-0.145	0.170	-0.196	0.104	-0.126	0.406
	a	45	-0.040	0.402	-0.011	0.473	0.030	0.851	-0.030	0.423	-0.012	0.467	0.042	0.734
	b	45	-0.090	0.289	-0.178	0.132	-0.323	0.035 *	-0.103	0.252	-0.231	0.060	-0.313	0.009 **
	c	45	-0.150	0.174	-0.230	0.073	-0.239	0.129	-0.086	0.289	-0.182	0.123	-0.167	0.239
	d	43	-0.128	0.219	-0.244	0.066	-0.159	0.334	-0.113	0.241	-0.185	0.124	-0.301	0.048 *
	f+i	44	-0.100	0.271	-0.048	0.385	0.054	0.743	-0.158	0.159	-0.119	0.336	0.013	0.919
	g	25	-0.203	0.189	-0.264	0.123	-0.045	0.848	-0.231	0.145	-0.336	0.049 *	-0.030	0.890
	h	44	-0.297	0.029 *	-0.237	0.069	-0.072	0.660	-0.214	0.075	-0.226	0.069	-0.090	0.550

Association between pathway position and ω, d_N and d_S values after controlling for expression level and breadth, codon bias, protein length and connectivity. *, $P < 0.05$; **, $P < 0.01$.

[a]Branch codes according to figure 1.

[b]One-tailed *P*-values.

[c]Two-tailed *P*-values.

[d]Using overall ω, d_N and d_S values.

[e]Excluding *O. anatinus* from the analysis.

Table S15. Partial correlation and path analysis (dataset 3)

			Partial correlation analysis						Path analysis					
			ω		d_N		d_S		ω		d_N		d_S	
# species	Branch[a]	*n*	ρ	P[b]	ρ	P[b]	ρ	P[c]	β	P[b]	β	P[b]	β	P[c]
6	all[d]	20	-0.214	0.215	-0.314	0.117	-0.089	0.747	-0.125	0.267	-0.198	0.167	-0.147	0.481
	a	20	-0.028	0.459	-0.060	0.414	0.164	0.550	-0.022	0.462	-0.025	0.458	0.008	0.967
	b	20	-0.343	0.094	-0.382	0.068	-0.371	0.150	-0.064	0.371	-0.200	0.140	-0.316	0.096
	c	20	-0.175	0.260	-0.256	0.170	-0.299	0.258	-0.024	0.456	-0.091	0.339	-0.322	0.133
	d	20	-0.323	0.109	-0.321	0.111	-0.080	0.773	-0.102	0.319	-0.093	0.301	-0.079	0.697
	e	20	-0.446	0.036 *	-0.248	0.178	0.150	0.583	-0.219	0.144	-0.215	0.156	0.236	0.242
	f	20	-0.327	0.106	-0.158	0.282	0.069	0.804	-0.138	0.263	-0.047	0.415	0.123	0.590
	g	18	-0.372	0.092	-0.391	0.079	0.155	0.603	-0.370	0.026 *	-0.348	0.046 *	0.010	0.964
	h	20	-0.502	0.018 *	-0.664	0.001 ***	-0.370	0.151	-0.207	0.123	-0.405	0.009 **	-0.283	0.083
	i	17	-0.147	0.319	-0.452	0.055	0.053	0.866	-0.276	0.047 *	-0.499	0.010 **	0.047	0.852
5[e]	all[d]	20	-0.310	0.120	-0.405	0.055	-0.194	0.477	-0.298	0.033 *	-0.322	0.042 *	-0.094	0.630
	a	20	-0.265	0.161	-0.280	0.147	0.103	0.709	-0.275	0.094	-0.244	0.128	0.100	0.592
	b	20	-0.506	0.017 *	-0.511	0.016 *	-0.217	0.424	-0.307	0.009 **	-0.403	0.002 **	-0.208	0.287
	c	20	-0.434	0.041 *	-0.481	0.024 *	-0.284	0.286	-0.250	0.089	-0.306	0.046 *	-0.194	0.350
	d	20	-0.496	0.020 *	-0.398	0.059	-0.188	0.491	-0.326	0.042 *	-0.212	0.121	-0.082	0.684
	f+i	20	-0.153	0.289	-0.199	0.232	-0.062	0.822	-0.316	0.051	-0.202	0.175	0.070	0.716
	g	18	-0.463	0.042 *	-0.497	0.029 *	-0.027	0.927	-0.394	0.014 *	-0.494	0.003 **	-0.220	0.359
	h	20	-0.426	0.045 *	-0.587	0.004 **	-0.243	0.366	-0.224	0.112	-0.336	0.023 *	-0.226	0.221

Association between pathway position and ω, d_N and d_S values after controlling for expression level and breadth, codon bias, protein length and connectivity. *, $P < 0.05$; **, $P < 0.01$; ***, $P < 0.001$.

[a]Branch codes according to figure 1.

[b]One-tailed *P*-values.

[c]Two-tailed *P*-values.

[d]Using overall ω, d_N and d_S values.

[e]Excluding *O. anatinus* from the analysis.

Table S16. Inputs and outputs of the insulin/TOR pathway.

		Inputs		Outputs		
Gene	node#[a]	Activatory	Inhibitory	Activatory	Inhibitory	Undirected
INSR	365	2	2	5	0	1
IGF1R	347	1	0	9	0	2
INSRR	—	0	0	0	0	0
IRS1	370	3	2	3	0	3
IRS2	881	2	1	0	0	0
AC069281.1	—	0	0	0	0	0
IRS4	—	0	0	0	0	0
DOK1	815	5	0	1	0	15
DOK2	—	0	0	0	0	0
DOK3	—	0	0	0	0	0
DOK4	—	0	0	0	0	0
DOK5	—	0	0	0	0	0
DOK6	—	0	0	0	0	0
FRS2	1196	1	2	2	0	0
FRS3	—	0	0	0	0	0
PIK3R1	468	5	0	1	0	0
PIK3R2	1272	0	0	0	0	10
PIK3R3	1269	0	0	0	0	2
PIK3CA	515	21	3	14	1	8
PIK3CB	1266	2	0	1	1	2
PIK3CG	516	2	0	2	0	0
PIK3CD	—	0	0	0	0	0
VEPH1	—	0	0	0	0	0
PDPK1	975	1	0	3	0	0
AKT1	22	20	6	24	5	2
AKT2	—	0	0	0	0	0
AKT3	1130	3	0	3	3	0
PRKCA	523	5	0	15	17	5
PRKCB	460	0	0	2	1	0
PRKCG	1531	0	0	3	0	1
PRKCD	1530	0	0	3	0	1
PRKCE	524	1	0	0	0	1
PRKCZ	1282	0	0	1	0	1
PRKCH	—	0	0	0	0	0
PRKCQ	—	0	0	0	0	0
PRKCI	1532	0	0	0	0	1
TSC1	688	1	0	0	1	0
FOXO1	843	1	2	0	0	2
FOXO3	249	3	1	2	0	1
FOXO4	17	0	0	1	0	0
FOXO6	—	0	0	0	0	0
GSK3A	307	3	5	5	6	2
GSK3B	308	1	2	3	3	0

TSC2	689	0	0	0	1	0
EIF2B5	—	0	0	0	0	0
GYS1	—	0	0	0	0	0
GYS2	—	0	0	0	0	0
MYC	141	0	0	2	0	2
MYCL1	—	0	0	0	0	0
MYCN	—	0	0	0	0	0
RHEB	1022	0	0	1	0	0
RHEBL1	—	0	0	0	0	0
MTOR	440	0	0	1	1	0
EIF4EBP1	201	2	1	1	0	0
EIF4EBP2	—	0	0	0	0	0
EIF4EBP3	844	0	0	0	0	0
RPS6KB1	488	4	0	3	2	0
RPS6KB2	—	0	0	0	0	0
EIF4E	208	3	2	0	0	5
EIF4E1B	—	0	0	0	0	0
EIF4E2	—	0	0	0	0	0
EIF4E3	—	0	0	0	0	0
RPS6	590	2	0	0	0	0
CYTH1	—	0	0	0	0	0
CYTH2	749	2	0	1	0	0
CYTH3	1205	0	0	1	0	0
CYTH4	—	0	0	0	0	0
PTEN	551	4	1	2	2	0

[a]When available, the node number in the dataset of Cui et al. (2007) is provided.

Table S17. Similarity in ω, d_N and d_S values among genes encoding interacting proteins after controlling for the relationship of these parameters with pathway position.

		ω		d_N		d_S	
Dataset	n [a]	X	P	X	P	X	P
2	33	0.024	0.004 **	0.076	0.005 **	1.804	0.835
	20[b]	0.030	0.070	0.094	0.124	2.292	0.987
3	33	0.023	0.011 *	0.082	0.029 *	1.259	0.608
	20[b]	0.028	0.079	0.093	0.127	1.605	0.940

*, $P < 0.05$; **, $P < 0.01$.
[a]Number of interactions considered.
[b]Analysis considering only the interactions used in *Drosophila* (Alvarez-Ponce et al. 2009).

REFERENCES

Alvarez-Ponce D, Aguadé M, Rozas J. 2009. Network-level molecular evolutionary analysis of the insulin/TOR signal transduction pathway across 12 Drosophila genomes. Genome Res. 19:234-242.

Cui Q et al. 2007. A map of human cancer signaling. Mol Syst Biol. 3:152.

Eyre TA et al. 2006. The HUGO Gene Nomenclature Database, 2006 updates. Nucleic Acids Res. 34:D319-321.

Karro JE et al. 2007. Pseudogene.org: a comprehensive database and comparison platform for pseudogene annotation. Nucleic Acids Res. 35:D55-60.

Su AI et al. 2004. A gene atlas of the mouse and human protein-encoding transcriptomes. Proc Natl Acad Sci U S A. 101:6062-6067.

Wright F. 1990. The 'effective number of codons' used in a gene. Gene. 87:23-29.

4. Discussió

En els treballs que constitueixen aquesta tesi s'han emprat tècniques de genòmica comparada i bioinformàtica, tant automàtiques com manuals, per tal d'identificar i anotar els gens implicats en la via de la insulina/TOR en els genomes de 12 espècies de *Drosophila* i 6 de vertebrats. Posteriorment, es van emprar tècniques d'anàlisi filogenètica per caracteritzar-ne els patrons d'evolució molecular, i es va investigar la seva possible relació amb la posició que ocupen els elements a la via.

A *Drosophila*, aquesta anàlisi s'ha realitzat sobre un total de 27 gens de *Drosophila melanogaster*, 7 dels quals són paràlegs que codifiquen els factors d'iniciació de l'elongació 4E (taula S1, article 1). A l'anàlisi de la via de la insulina/TOR dels vertebrats, el nombre de gens humans estudiats és de 115 (72 gens potencialment funcionals i 43 pseudogens; taula S1, article 2). La diferència en el nombre de gens emprats en ambdós estudis rau principalment en el fet que en els genomes de vertebrats la majoria dels gens estudiats existeixen en múltiples còpies. De fet, els 115 gens humans estudiats pertanyen a 24 grups de paràlegs. En l'estudi de la via de la insulina/TOR a vertebrats, s'han emprat tres conjunts de dades: el conjunt de dades 1 inclou tots els gens estudiats; el conjunt de dades 2 n'inclou un subconjunt, amb només un gen per grup de paràlegs, triat d'acord amb el coneixement sobre la funció molecular de les diferents còpies de cada grup; i el conjunt de dades 3 s'ha obtingut mitjançant el càlcul, per a cada grup de paràlegs, de la mitjana dels valors dels diferents gens.

4.1. Identificació dels gens implicats en la via de transducció de senyal de la insulina/TOR en els genomes de *Drosophila* i vertebrats

En total s'han identificat 342 gens i pseudogens a les 12 espècies de *Drosophila* estudiades (335 gens potencialment funcionals i 7 pseudogens; taula S2, article 1) i 734 a les 6 espècies de vertebrats (400 gens potencialment funcionals, 43 còpies sense introns i 289 pseudogens; taula S5, article 2). Aquests nombres, però, s'han de considerar com a estimes mínimes dels nombres reals, ja que (1) tots els genomes estudiats tenen regions no

seqüenciades; i (2) els gens duplicats recentment, que s'espera que siguin força similars, podrien haver estat erròniament tractats com una única còpia durant el procés d'ensamblat. No obstant això, cal destacar que per a totes les espècies estudiades es disposa de seqüències genòmiques amb una elevada cobertura (entre 6X i 10X), i per tant és altament probable que aquests genomes realment no pressentin els gens no trobats en les nostres anàlisis.

A *Drosophila*, tots els gens de la via de la insulina/TOR tenen almenys un ortòleg potencialment funcional en totes les espècies estudiades, amb excepció del gen *eIF4E-6*, que només està present en les espècies del subgrup *melanogaster* (taula S2, article 1). Aquest gen pertany a un grup de paràlegs, amb 7 membres a *D. melanogaster* (Hernandez et al. 2005), i, a l'igual que el gen *4EHP*, podria correspondre a una còpia no funcional o inclús un regulador negatiu de la resta de eIF4Es (Hernandez et al. 2005). Per tant, les nostres observacions indiquen que els 12 genomes de *Drosophila* estudiats codificarien una via de la insulina/TOR completa.

En els vertebrats, tot i que alguns dels gens estudiats no presenten còpies potencialment funcionals en tots els genomes estudiats (12 dels 72 gens estudiats), tots els grups de paràlegs presenten almenys una còpia no pseudogènica en les sis espècies estudiades (taula S5, article 2). Per tant, la funció dels gens no presents en alguna de les espècies probablement sigui suplida per algun dels seus paràlegs funcionals, amb la qual cosa també les espècies de vertebrats estudiades presentarien una via completa de la insulina/TOR.

A *Drosophila*, les dades del nombre de còpies de cada gen a cada genoma, combinades amb una anàlisi de reconstrucció filogenètica, van permetre inferir que s'ha donat un total de 20 esdeveniments de duplicació gènica, un de pèrdua de gens i 5 de pseudogenització, així com en quines branques de la filogènia haurien tingut lloc aquests esdeveniments (figura 1, article 1).

4.2. Impacte de la selecció natural sobre els gens de la via de la insulina/TOR

S'ha caracteritzat l'impacte de la selecció natural sobre l'evolució molecular dels gens de la via de la insulina/TOR a partir del grau de divergència no sinònima (d_N), i de la relació entre aquest i la divergència sinònima ($\omega = d_N/d_S$). Totes les estimes de ω obtingudes estan clarament per sota de 1 [amb valors màxims de 0,220 a *Drosophila* (taula 1, article 1) i 0,140 a vertebrats (taula S3, article 2)], la qual cosa indica que els gens de la via de la insulina/TOR han evolucionat sota una relativament forta selecció purificadora. Això, juntament amb el fet que tots els genomes estudiats presenten almenys un representant potencialment funcional de tots els grups de paràlegs de la via de la insulina/TOR (veure apartat 4.1), indica que totes les espècies estudiades tenen una via de la insulina/TOR completa i funcional.

Es van aplicar dos tests estadístics per detectar la petjada de la selecció positiva sobre la relació entre la divergència no sinònima i la sinònima (la presència de codons amb $\omega > 1$): els tests M1a *vs.* M2a (Wong et al. 2004), i el test M7 *vs.* M8 (Yang et al. 2000). Un ajust significativament millor dels models M2a o M8 a les dades s'atribueix a l'efecte de la selecció positiva. A *Drosophila*, tot i que el test M7 *vs.* M8 és significatiu per als gens *eIF2B-ε*, *Akt1* i *Tor*, quan s'aplica una correcció per a tests múltiples (*false discovery rate*; Benjamini i Hochberg 1995), el test només és significatiu per als gens *eIF2B-ε* i *Akt1* (taula S3, article 1). Als vertebrats, el test M7 *vs.* M8 mostra que els gens *IRS4*, *AKT3* i *PRKCD* han evolucionat sota selecció positiva, tot i que cap dels resultats és significatiu quan es controla per a tests múltiples (taula S6, article 2). La selecció positiva no semblaria haver estat un factor rellevant en l'evolució molecular dels gens de la via de la insulina/TOR en cap dels dos llinatges.

4.3. Relació entre l'estructura de la via de la insulina/TOR i els patrons d'evolució molecular dels seus components

Un cop caracteritzats els patrons d'evolució molecular dels gens implicats en la via de transducció de senyal de la insulina/TOR a *Drosophila* i als vertebrats, es va investigar la possible relació entre l'estructura de la xarxa d'interaccions que connecta els elements de la via i els patrons d'evolució molecular dels seus components.

Per tal de dur a terme aquesta tasca, es va codificar l'estructura de la via de la insulina en forma de graf dirigit (anomenat *G*), en el qual els nodes representen les proteïnes de la via, i els arcs representen les interaccions entre aquestes (tant d'activació com d'inhibició). A *Drosophila*, el graf consisteix en 19 nodes connectats per 25 arcs (figura S2A, article 1), mentre que a vertebrats consisteix en 21 nodes connectats per 39 arcs (figura 2A). La diferència en el nombre d'arcs rau majoritàriament en el fet que s'ha identificat un major nombre d'interaccions a vertebrats.

Aquest graf va ser emprat per investigar (1) si els gens que codifiquen proteïnes que interactuen físicament tendeixen a evolucionar de manera semblant (apartat 4.3.1); i (2) si existeix una polaritat en els nivells de limitació funcional a variar al llarg de l'eix *upstream*/*downstream* de la via (apartat 4.3.2).

4.3.1. Similitud en els patrons d'evolució molecular dels gens que codifiquen proteïnes que interactuen físicament

Es va emprar el mètode de Monte Carlo descrit per Fraser i col·laboradors (2002) per tal de determinar si els gens que codifiquen proteïnes que interactuen físicament tendeixen a evolucionar sota una selecció purificadora d'intensitat semblant. En aquesta anàlisi es va emprar un subgraf del graf esmentat anteriorment (denominat *S*) que conté únicament les interaccions físiques proteïna–proteïna (PPIs): 20 interaccions a *Drosophila* i 32 a vertebrats (figura 2C, article 1; figura 2B, article 2). Aquesta anàlisi mostra que, tant a

Drosophila com a vertebrats, els gens que codifiquen proteïnes que interactuen físicament presenten nivells de limitació funcional semblants. L'aplicació del mateix tipus d'anàlisi a les dues components de ω (d_N i d_S), dóna resultats significatius per a d_N però no per a d_S, la qual cosa indica que l'evolució a nivell de seqüència aminoacídica seria la principal responsable de la tendència observada. Aquesta tendència, també observada a escala interactòmica, s'ha atribuït a la coevolució dels aminoàcids implicats en les interaccions proteïna-proteïna, o també a que les proteïnes que interactuen estarien sotmeses a nivells similars de selecció estabilitzadora (Fraser et al. 2002; Lemos et al. 2005). No obstant això, en la via de la insulina/TOR aquest patró també podria ser un subproducte de la correlació entre els nivells de limitació funcional i la posició dels gens a la via (veure apartat 4.3.2): Atès que les proteïnes de la via tendeixen a interactuar amb d'altres que ocupen posicions adjacents, una correlació entre la posició dels elements a la via i el nivell de limitació funcional podria resultar en una similitud en els nivells de limitació funcional de les proteïnes que interactuen entre sí. Per tal de descartar aquesta possibilitat, es van repetir les anàlisis tot descomptant l'efecte de l'associació entre la posició i el nivell de limitació funcional. Els resultats són diferents a *Drosophila* i vertebrats: mentre que a *Drosophila* la tendència desapareix, a vertebrats continua sent significativa. Aquesta diferència podria ser deguda al diferent nombre d'interaccions emprades en ambdues anàlisis: atès que a *Drosophila* només se n'utilitzen 20 (un subconjunt de les 32 emprades a vertebrats), això podria comportar una reducció en la potència estadística de les anàlisis realitzades a *Drosophila*. De manera consistent amb aquesta possibilitat, quan l'anàlisi de la via de la insulina/TOR a vertebrats es restringeix a les 20 interaccions considerades a *Drosophila*, s'obtenen resultats equivalents als obtinguts a *Drosophila*: si bé les proteïnes que interactuen físicament tendeixen a presentar nivells semblants de ω (taula S7, article 2), aquesta tendència desapareix quan es descompta l'efecte de la correlació entre ω i la posició dels elements a la via (taula S17, article 2).

Les nostres anàlisis, per tant, indiquen que les proteïnes que interactuen entre si evolucionen sota uns nivells de limitació funcional semblants, probablement degut a la coevolució dels llocs implicats en interaccions proteïna-

proteïna o al fet que aquestes proteïnes estarien sotmeses a nivells semblants de selecció estabilitzadora.

S'ha suggerit que els gens que codifiquen proteïnes que interactuen entre si tendirien a presentar patrons de duplicació i pèrdua semblants, degut a una evolució coordinada (Fryxell 1996). Es va analitzar aquesta possible tendència (duplicació als mateixos llinatges) en els gens de la via de la insulina/TOR de *Drosophila*, i es va veure que els gens que codifiquen 4 proteïnes que interactuen físicament (*eIF-4E*, *Akt1*, *Thor* i *Tor*) s'han duplicat al llinatge de *D. willistoni*, i que 2 gens que també codifiquen proteïnes que interactuen entre si (*eIF4E-3* i *Thor*) s'han duplicat al llinatge de *D. grimshawi* (figura 1, article 1). Per tal de contrastar si aquestes observacions són compatibles amb una generació aleatòria de duplicacions al llarg de l'evolució, es va emprar un mètode de Monte Carlo. Aquesta anàlisi no permet rebutjar la hipòtesi nul·la de que els esdeveniments de duplicació haurien tingut lloc de manera aleatòria al llarg de la filogènia.

4.3.2. Polaritat en els nivells de limitació funcional al llarg de l'eix *upstream/downstream* de la via

Es va considerar la possible relació entre el nivell de limitació funcional a variar que actua sobre els gens de la via de la insulina/TOR i la posició que aquests ocupen al llarg de l'eix *upstream*/*downstream* de la mateixa. A aquest efecte, es va empra el graf *G* per assignar a cada element de la via una posició, comptada com el nombre de passos necessaris per a la transducció del senyal des del receptor de la insulina (que ocupa la posició 0) fins la resta d'elements de la via (el màxim nombre de passos és de 10; figura 2, article 1; figura 2, article 2).

A *Drosophila*, es va trobar una correlació significativa entre la posició dels gens a la via i ω, essent els gens de la part *downstream* els que evolucionen sota pressions selectives més fortes. De la mateixa manera que ω, d_N es correlaciona amb la posició dels elements a la via, mentre que d_S no ho fa, la qual cosa apunta a la variabilitat en les taxes de divergència no sinònima com a principal responsable de la tendència observada. Es va avaluar la correlació entre la

posició a la via i ω de manera separada per a les 9 branques de la filogènia de les 6 espècies de *Drosophila* emprades per les nostres anàlisis de divergència (figura 1, article 1). La correlació és significativa i amb signe negatiu per a 3 de les branques i, tot i que no és significativa per a les 6 branques restants, (1) el sentit de la correlació és negatiu per a 5 d'aquestes branques; i (2) quan es va emprar un model específic que considera una única ω per a aquestes 6 branques, les estimes de ω obtingudes es correlacionen significativament amb la posició a la via. Aquestes observacions indiquen que la polaritat observada en els nivells de limitació funcional al llarg de la via no resultaria dels patrons d'evolució d'una branca en particular, sinó que seria una tendència més general.

Si bé als vertebrats la correlació entre la posició dels gens a la via i els valors globals de limitació funcional no és significativa (taules 2, S8 i S9, article 2), (1) el sentit de la correlació és sempre negatiu, independentment de la mesura de limitació funcional (ω o d_N) i del conjunt de dades emprats (taules 2, S8 i S9, article 2); i (2) quan s'elimina de les anàlisis el genoma de l'ornitorinc (ja que la seqüència disponible en l'actualitat està altament fragmentada), la correlació entre d_N i la posició a la via resulta significativa per al conjunt de dades 2 (taules 2, S8 i S9, article 2). A més, quan s'analitza la correlació separadament per a les 9 branques de la filogènia, es troba que la correlació entre la posició a la via i ω o d_N és negativa per a la majoria de les branques (una proporció generalment superior al 50% que s'esperaria si no hi hagués una polaritat en els nivells de limitació funcional al llarg de la via; taules 3, S10 i S11, article 2). Aquests resultats indiquen que també als vertebrats els nivells de limitació funcional seguirien un gradient, amb nivells més elevats als gens de la part *downstream*. Per tant, els mateixos mecanismes biològics implicats a la polaritat en els nivells de limitació funcional al llarg de la via haurien estat operant sobre la via de la insulina d'ambdós llinatges. No obstant això, la correlació seria menys clara als vertebrats (discutit a l'apartat 4.4).

Aquestes observacions indiquen que l'evolució dels gens de la via de la insulina/TOR està limitada per la posició que ocupen a la via, i per tant que l'estructura de la via té un efecte sobre l'evolució molecular dels seus components. Si bé en altres vies també s'ha observat una correlació entre els nivells de limitació funcional dels gens i la posició que ocupen, incloent-hi la

ruta de les antocianines, la de l'isoprè, la dels terpenoides i la dels carotenoides (Rausher et al. 1999; Sharkey et al. 2005; Livingstone i Anderson 2009; Ramsay et al. 2009), i també la via de senyalització Ras de *Drosophila* (Riley et al. 2003), el sentit de la correlació és sempre negatiu, amb un major impacte de la selecció purificadora als gens que actuen a la part *upstream*, en contraposició amb el sentit de la correlació observada a la via de la insulina/TOR.

Es van considerar una sèrie de possibles explicacions per a aquesta tendència. De fet, hi ha tota una sèrie de factors que es correlacionen amb ω i d_N. Per tant, una possible polaritat en qualsevol d'aquests factors al llarg de la via podria donar compte, totalment o parcialment, de la polaritat en els nivells de limitació funcional observats. Així la selecció positiva incrementaria el nivell de divergència no sinònima, amb la qual cosa si la intensitat de la selecció positiva presentés un gradient al llarg de la via, amb un major impacte als gens de la part *upstream*, aquest gradient podria explicar la correlació observada entre la posició a la via i ω i d_N. No obstant això, (1) com ja s'ha esmentat anteriorment, la selecció positiva no sembla haver tingut un impacte important sobre l'evolució dels gens de la via; i (2) ni a *Drosophila* ni a vertebrats s'ha trobat una distribució particular dels gens que presenten algun senyal d'haver evolucionat de manera adaptativa (a *Drosophila*, aquests gens ocupen les posicions 5, 7 i 8, i a vertebrats els gens amb la petjada de la selecció positiva ocupen les posicions 1, 5 i 6). Per tant, la selecció positiva podria ser descartada com a factor responsable de la polaritat en els valors de ω i d_N.

D'altra banda, s'ha observat que els gens amb elevats nivells d'expressió, amb una elevada amplitud d'expressió, amb un ús fortament esbiaixat de codons, amb uns elevats nivells de connectivitat, o que codifiquen proteïnes més curtes, tendeixen a presentar nivells més elevats de limitació funcional (Sharp 1991; Duret i Mouchiroud 2000; Pál et al. 2001; Fraser et al. 2002; Subramanian i Kumar 2004), i per tant evolucionarien més lentament. Es va avaluar, per tant, la possible implicació d'aquests factors en la polaritat en els nivells de limitació funcional al llarg de la via. A aquest efecte, es va avaluar si algun d'aquest factors es correlaciona amb la posició a la via i/o amb els valors de ω i d_N. A més, es van emprar dues tècniques d'anàlisi multivariant, l'anàlisi de camins i la

correlació parcial, per tal d'avaluar l'associació entre la posició a la via i ω i d_N un cop descomptat l'efecte de tots aquests factors.

A *Drosophila*, si bé el nivell d'expressió i el biaix en l'ús de codons presenten una correlació positiva amb la posició a la via, i la longitud de les proteïnes hi presenta una correlació negativa (figura S2, article 1), cap dels factors estudiats es correlaciona amb d_N ni amb ω (figura S2, article 1), la qual cosa suggereix que aquests factors no serien els responsables de la polaritat en els nivells de limitació funcional al llarg de la via. Tant l'anàlisi de camins (figura 4, article 1) com la de correlacions parcials[10] confirmen que l'associació entre els nivells de limitació funcional i la posició dels gens a la via és independent d'aquests factors.

A vertebrats, quan s'apliquen les tècniques d'anàlisi multivariant s'obtenen resultats idèntics als que s'obtenen de l'anàlisi de correlació bivariada: (1) el sentit de l'associació entre la posició a la via i ω (o d_N) és sempre negatiu, independentment del conjunt de dades emprat (taula S12, article 2), i (2) el sentit d'aquesta associació és negatiu en un nombre de branques de la filogènia significativament superior al 50% (taules 4, S14 i S15, article 2). A més, l'anàlisi de camins mostra que l'associació entre d_N i la posició a la via és significativa per al conjunt de dades 2. Aquests cinc factors, per tant, tampoc serien els responsables de la polaritat en els nivells de limitació funcional observada al llarg de la via de la insulina/TOR dels vertebrats.

Un altre factor que pot influenciar les pressions selectives que actuen sobre un gen és el nombre de vies en les quals està implicat. Així, s'espera que les mutacions que tenen lloc en els gens que participen en un major nombre de vies tinguin efectes pleiotròpics més importants, i per tant que aquests gens estiguin sotmesos a un major grau de limitació funcional. El nombre de vies en què estan implicats els gens que actuen en les diferents parts d'una via depèn de la manera en què aquesta està connectada amb altres vies. Així, en una via lineal que rep múltiples entrades (ja sigui d'informació, en el cas d'una via de transducció de senyal, o metabòlits, en el cas d'una via metabòlica) i amb una

[10]Anàlisi no inclosa al primer article. Coeficient de correlació parcial entre ω i la posició a la via, $\rho = -0.573$, $P = 0.025$.

única sortida, els gens que actuen a la part *downstream* estarien implicats en un major nombre de vies, i per tant presumiblement més limitats a variar. Per contra, en una via amb una única entrada que emet informació (o metabòlits) cap a altres vies, s'espera un major grau de limitació funcional a la part *upstream* (figura 5, article 1).

El sentit de la correlació entre el grau de limitació funcional dels gens de la ruta de biosíntesi de les antocianines i la posició que ocupen a la via (amb un major impacte de la selecció purificadora en els gens que actuen a la part *upstream*) concorda amb aquest model. Degut a l'estructura ramificada d'aquesta via, els gens situats a la part *upstream* estan implicats en la síntesi de tota una sèrie de compostos, mentre que l'enzim que catalitza el darrer pas de la via està implicat únicament en la síntesi de les antocianines (figura 3). Aquesta polaritat en el nombre de compostos en la síntesi dels quals està implicat cadascun d'aquests enzims podria explicar la distribució de la limitació funcional al llarg de l'eix *upstream*/*downstream* de la via de les antocianines (Rausher et al. 1999). Les altres rutes biosintètiques en les quals s'ha observat una correlació entre la posició a la via i els nivells de limitació funcional presenten estructures similars, amb ramificacions que determinen la implicació dels elements de la part *upstream* en la biosíntesi d'un major nombre de compostos.

D'acord amb aquest model, la polaritat observada en els nivells de limitació al llarg de la via de la insulina/TOR podria ser deguda a una possible implicació dels elements de la part *downstream* en un major nombre de vies que els elements de la part *upstream*. Aquesta distribució del nombre de vies en què estan implicats els elements que actuen en les diferents parts de la via s'esperaria si el nombre d'entrades d'informació superés el nombre de sortides. Per tal de contrastar aquesta possibilitat, es van emprar diferents tipus d'informació sobre els patrons de connexió de la via de la insulina amb altres vies. A *Drosophila*, es va cercar aquesta informació a la literatura (taula S4, article 1), mentre que a vertebrats es va emprar un conjunt de dades generat per Cui i col·laboradors (2007), que consisteix en 5089 interaccions dirigides entre 1634 proteïnes humanes, en 356 de les quals participa únicament una proteïna de la via de la insulina/TOR (taula S16, article 2). En cap dels dos

casos es fa evident un major nombre d'entrades que de sortides, la qual cosa no recolzaria els patrons de connexió de la via de la insulina amb altres vies com a factor responsable de la polaritat observada en els nivells de limitació funcional. Cal destacar, però, que (1) el coneixement actual de la manera en què la via de la insulina/TOR està connectada amb altres vies és força limitat; i (2) la relació entre el nombre d'entrades i sortides en què es troben implicats els elements d'una via probablement sigui un predictor poc acurat de la distribució dels nivells de limitació funcional al llarg de l'eix *upstream*/*downstream*. Atès que les diferents interaccions diferirien en quant a rellevància (per exemple, una molècula o via pot activar de manera molt important la via de la insulina, mentre que una altra podria fer-ho amb molt poca intensitat), un predictor més adient seria la rellevància relativa, en termes d'efectes sobre l'eficàcia biològica, de les entrades respecte de les sortides. El coneixement actual sobre els patrons de connexió de la via de la insulina, però, dista molt de permetre obtenir aquest tipus de mesures amb precisió. Per tant, és prematur valorar l'efecte dels patrons de connexió de la via de la insulina/TOR amb altres vies sobre els patrons d'evolució dels seus components.

4.3.3. Evolució dels gens que actuen als punts de ramificació de la via de la insulina/TOR

Les proteïnes d'una via, ja sigui metabòlica o de transducció de senyal, poden contribuir de manera diferencial a la funció d'aquesta. S'espera que els gens la variabilitat dels quals té un efecte més important sobre la funció global de la via (i, per tant, sobre els fenotips en els quals aquesta està implicada) evolucionin sota pressions selectives més fortes que els gens que tenen un efecte menys important sobre la funció de la via (Hartl et al. 1985; Eanes 1999; Watt i Dean 2000). Per tant, una anàlisi de la sensibilitat de la funció de la via a les propietats de cadascun dels seus components podria ajudar a explicar la distribució dels nivells de limitació funcional. Recentment s'ha desenvolupat un model matemàtic per a aquesta via (Zielinski et al. 2009), que podria servir com a punt de partida per a una anàlisi d'aquest tipus.

Els enzims que actuen en els punts de ramificació d'una via jugarien un paper clau en el control del flux al llarg de les diferents branques d'aquesta (LaPorte et al. 1984; Stephanopoulos i Vallino 1991), i per tant s'espera que siguin, de manera preferent, la diana de la selecció natural. De manera consistent amb aquesta hipòtesi, 5 gens que actuen en punts de ramificació de les vies del catabolisme de la glucosa a *Drosophila* presenten la petjada de la selecció positiva (figura 4; Flowers et al. 2007). Pel que fa als gens de la via de la insulina/TOR, a *Drosophila* dos dels tres gens que presenten alguna evidència d'haver evolucionat de manera adaptativa (*Akt1* i *Tor*) codifiquen enzims que actuen en punts de ramificació de la via. A vertebrats, també dos dels tres gens que presenten codons amb $\omega > 1$ (*AKT3* i *PRKCD*) actuen en punts de ramificació de la via.

4.4. Comparació dels patrons d'evolució molecular de la via de la insulina a *Drosophila* i a vertebrats

Tot i que tant a *Drosophila* com a vertebrats s'observa una polaritat en els nivells de limitació funcional al llarg de l'eix *upstream*/*downstream* de la via de la insulina/TOR, aquesta polaritat és menys clara als vertebrats. Aquesta diferència podria deure's a diferents factors. En primer lloc, atès que els genomes de vertebrats presenten taxes de substitució molt inferiors als de *Drosophila* (Sharp i Li 1989), les anàlisis realitzades als vertebrats podrien estar basades en un menor nombre de substitucions nucleotídiques, amb la conseqüent reducció en la potència estadística. No obstant això, la comparació dels nivells de divergència sinònima i del nombre total de canvis no sinònims que haurien tingut lloc mostren que, de fet, tots dos paràmetres presenten valors superiors als vertebrats. Això es pot explicar perquè les espècies de vertebrats estudiades van divergir fa uns 310 milions d'anys (figura 8), en contraposició amb els aproximadament 44 milions d'anys transcorreguts des de la separació de les 6 espècies del grup *melanogaster* emprades a l'anàlisi de l'impacte de la selecció natural (Tamura et al. 2004; figura 6).

En segon lloc, els vertebrats presenten en general una grandària efectiva més petita que les espècies de *Drosophila* (Lynch i Conery 2003). La teoria quasi

neutralista de l'evolució molecular prediu una menor eficiència de la selecció natural en poblacions amb una reduïda grandària efectiva, ja que el destí de les mutacions lleugerament deletèries estaria determinat, en gran mesura, per la deriva genètica (Ohta 1973). De manera consistent, s'ha vist que la selecció purificadora és més forta a *Drosophila* que a vertebrats (veure, per exemple, Petit i Barbadilla 2009). Per tant, els mecanismes biològics que mantindrien la polaritat en els nivells de limitació funcional al llarg de la via de la insulina/TOR serien menys eficients als vertebrats, la qual cosa justificaria la menor claretat de la polaritat observada. No obstant això, als gens de la via de la insulina/TOR no s'observa una diferència en els valors de ω entre ambdós llinatges.

Per acabar, tot i que a *Drosophila* la major part dels gens de la via de la insulina/TOR són de còpia única (taula S1 i S2, article 1), als genomes dels vertebrats la majoria dels gens de la via existeixen en múltiples còpies (taula S1, article 2). Atès que el nivell de limitació funcional d'un gen depèn del nombre de còpies que presenta (Lynch i Conery 2000; Jordan et al. 2004), la polaritat en els nivells de limitació funcional al llarg de la via de la insulina/TOR podria ser el resultat d'una possible polaritat en el nombre de còpies dels seus gens. No obstant això, als gens de la via de la insulina/TOR de vertebrats el nombre de còpies de cada grup de paràlegs no es correlaciona ni amb la posició a la via ni amb els valors mitjans de ω ni de d_N, amb la qual cosa aquest factor podria ésser descartat com a responsable de la distribució de l'impacte de la selecció purificadora al llarg de la via.

5. Conclusions

1. Tots els genomes estudiats presenten almenys una còpia potencialment funcional de tots els grups de gens estudiats. Per tant, totes les espècies estudiades codificarien una via de la insulina/TOR completa.

2. Les estimes de ω obtingudes oscil·len entre 0,002 i 0,220, cosa que indica que tots els gens estudiats han evolucionat sota una relativament forta selecció purificadora. Tots aquests gens, per tant, serien funcionals, amb la qual cosa els genomes estudiats codificarien una via de la insulina/TOR funcional. A més, els gens *eIF2B-ε*, *Akt1* i *Tor* de *Drosophila* i els gens *IRS4*, *AKT3* i *PRKCD* de vertebrats presenten la petjada de la selecció positiva. No obstant això, només els gens *eIF2B-ε* i *Akt1* de *Drosophila* donen resultats significatius quan s'aplica una correcció per múltiples tests.

3. A *Drosophila*, s'ha inferit un total de 20 esdeveniments de duplicació, un de pèrdua i 5 de pseudogenització en els gens de la via de la insulina/TOR

4. Tant a *Drosophila* com a vertebrats hi ha una polaritat en els nivells de limitació funcional al llarg de l'eix *upstream*/*downstream* de la via, essent els gens que actuen a la part *downstream* els que evolucionen sota una selecció purificadora més forta. Aquesta tendència és independent de la distribució de diferents factors que afecten la taxa d'evolució (selecció positiva, nivell i amplitud de l'expressió gènica, biaix en l'ús de codons, connectivitat i longitud de les proteïnes codificades). El sentit d'aquesta polaritat contrasta amb l'observat en altres vies, on els gens que actuen a la part *upstream* són els que evolucionen sota una selecció purificadora més intensa.

5. Els gens que codifiquen proteïnes que interactuen físicament han evolucionat sota nivells de limitació funcional semblants. A vertebrats, aquesta similitud és independent de la polaritat detectada en els nivells de limitació funcional al llarg de l'eix *upstream*/*downstream* de la via.

6. Quatre gens que codifiquen per proteïnes que interactuen físicament (*eIF-4E*, *Akt1*, *Thor* i *Tor*) s'han duplicat al llinatge de *D. willistoni*, mentre que 2 gens que també codifiquen proteïnes que interactuen entre si

(*eIF4E-3* i *Thor*) s'han duplicat al llinatge de *D. grimshawi*. Aquestes observacions, però, no són incompatibles amb una distribució aleatòria dels esdeveniments de duplicació al llarg de la filogènia de les 12 espècies estudiades.

7. Dos dels gens que presenten senyals de selecció positiva a *Drosophila* (*Akt1* i *Tor*), i dos a vertebrats (*AKT3* i *PRKCD*), actuen en punts de ramificació clau de la via. Això podria reflectir una tendència dels gens que actuen en aquestes parts de la via a evolucionar de manera adaptativa.

8. Globalment, els patrons d'evolució molecular i el destí evolutiu dels gens de la via de la insulina/TOR depenen de la posició particular que ocupen a la via.

6. Bibliografia

Adams MD et al. 2000. The genome sequence of Drosophila melanogaster. Science. 287:2185-2195.

Akashi H. 1994. Synonymous codon usage in Drosophila melanogaster: natural selection and translational accuracy. Genetics. 136:927-935.

Akashi H. 2003. Translational selection and yeast proteome evolution. Genetics. 164:1291-1303.

Akashi H. 2001. Gene expression and molecular evolution. Curr Opin Genet Dev. 11:660-666.

Albert R, Barabási AL. 2002. Statistical mechanics of complex networks. Rev. Mod. Phys. 74:47-97.

Avruch J, Belham C, Weng Q, Hara K, Yonezawa K. 2001. The p70 S6 kinase integrates nutrient and growth signals to control translational capacity. Prog Mol Subcell Biol. 26:115-154.

Banting FG, Best CH, Collip JB, Campbell WR, Fletcher AA. 1922. Pancreatic extracts in the treatment of diabetes mellitus: preliminary report. CMAJ. 2:141-146.

Benjamini Y, Hochberg Y. 1995. Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing. J. R. Statist. Soc. B. 57:289-300.

Betancourt AJ, Presgraves DC. 2002. Linkage limits the power of natural selection in Drosophila. Proc Natl Acad Sci U S A. 99:13616-13620.

Bossi A, Lehner B. 2009. Tissue specificity and the human protein interaction network. Mol Syst Biol. 5:260.

Brogiolo W et al. 2001. An evolutionarily conserved function of the Drosophila insulin receptor and insulin-like peptides in growth control. Curr Biol. 11:213-221.

Cai SL et al. 2006. Activity of TSC2 is inhibited by AKT-mediated phosphorylation and membrane partitioning. J Cell Biol. 173:279-289.

Castillo-Davis CI, Kondrashov FA, Hartl DL, Kulathinal RJ. 2004. The functional genomic distribution of protein divergence in two animal phyla: coevolution, genomic conflict, and constraint. Genome Res. 14:802-811.

Ceddia RB, Bikopoulos GJ, Hilliker AJ, Sweeney G. 2003. Insulin stimulates glucose metabolism via the pentose phosphate pathway in Drosophila Kc cells. FEBS Lett. 555:307-310.

Chou MM, Blenis J. 1995. The 70 kDa S6 kinase: regulation of a kinase with multiple roles in mitogenic signalling. Curr Opin Cell Biol. 7:806-814.

Comeron JM, Kreitman M. 1998. The correlation between synonymous and nonsynonymous substitutions in Drosophila: mutation, selection or relaxed constraints? Genetics. 150:767-775.

Cork JM, Purugganan MD. 2004. The evolution of molecular genetic pathways and networks. Bioessays. 26:479-484.

Cui Q et al. 2007. A map of human cancer signaling. Mol Syst Biol. 3:152.

Drosophila 12 Genomes Consortium. 2007. Evolution of genes and genomes on the Drosophila phylogeny. Nature. 450:203-218.

Drummond DA, Bloom JD, Adami C, Wilke CO, Arnold FH. 2005. Why highly expressed proteins evolve slowly. Proc Natl Acad Sci U S A. 102:14338-14343.

Drummond DA, Raval A, Wilke CO. 2006. A single determinant dominates the rate of yeast protein evolution. Mol Biol Evol. 23:327-337.

Dufner A, Thomas G. 1999. Ribosomal S6 kinase signaling and the control of translation. Exp Cell Res. 253:100-109.

Duret L, Mouchiroud D. 2000. Determinants of substitution rates in mammalian genes: expression pattern affects selection intensity but not mutation rate. Mol Biol Evol. 17:68-74.

Duret L, Mouchiroud D. 1999. Expression pattern and, surprisingly, gene length shape codon usage in Caenorhabditis, Drosophila, and Arabidopsis. Proc Natl Acad Sci U S A. 96:4482-4487.

Eanes WF. 1999. Analysis of selection on enzyme polymorphisms. Rev Ecol Syst. 30:301-326.

Fernandez R, Tabarini D, Azpiazu N, Frasch M, Schlessinger J. 1995. The Drosophila insulin receptor homolog: a gene essential for embryonic development encodes two receptor isoforms with different signaling potential. EMBO J. 14:3373-3384.

Flowers JM et al. 2007. Adaptive evolution of metabolic pathways in Drosophila. Mol Biol Evol. 24:1347-1354.

Franke TF. 2008. PI3K/Akt: getting it right matters. Oncogene. 27:6473-6488.

Fraser HB, Hirsh AE, Steinmetz LM, Scharfe C, Feldman MW. 2002. Evolutionary rate in the protein interaction network. Science. 296:750-752.

Fryxell KJ. 1996. The coevolution of gene family trees. Trends Genet. 12:364-369.

Galletti M et al. 2009. Identification of domains responsible for ubiquitin-dependent degradation of dMyc by glycogen synthase kinase 3beta and casein kinase 1 kinases. Mol Cell Biol. 29:3424-3434.

Giot L et al. 2003. A protein interaction map of Drosophila melanogaster. Science. 302:1727-1736.

Grewal SS, Li L, Orian A, Eisenman RN, Edgar BA. 2005. Myc-dependent regulation of ribosomal RNA synthesis during Drosophila development. Nat Cell Biol. 7:295-302.

Haddrill PR, Halligan DL, Tomaras D, Charlesworth B. 2007. Reduced efficacy of selection in regions of the Drosophila genome that lack crossing over. Genome Biol. 8:R18.

Hahn MW, Kern AD. 2005. Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks. Mol Biol Evol. 22:803-806.

Hartl DL, Dykhuizen DE, Dean AM. 1985. Limits of adaptation: the evolution of selective neutrality. Genetics. 111:655-674.

Hernandez G et al. 2005. Functional analysis of seven genes encoding eight translation initiation factor 4E (eIF4E) isoforms in Drosophila. Mech Dev. 122:529-543.

Hill WG, Robertson A. 1966. The effect of linkage on limits to artificial selection. Genet Res. 8:269-294.

Hirsh AE, Fraser HB. 2001. Protein dispensability and rate of evolution. Nature. 411:1046-1049.

Hubbard TJ et al. 2009. Ensembl 2009. Nucleic Acids Res. 37:D690-697.

Ingvarsson PK. 2007. Gene expression and protein length influence codon usage and rates of sequence evolution in Populus tremula. Mol Biol Evol. 24:836-844.

International Chicken Genome Sequencing Consortium. 2004. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. Nature. 432:695-716.

International Human Genome Sequencing Consortium. 2001. Initial sequencing and analysis of the human genome. Nature. 409:860-921.

International Human Genome Sequencing Consortium. 2004. Finishing the euchromatic sequence of the human genome. Nature. 431:931-945.

Jeong H, Tombor B, Albert R, Oltvai ZN, Barabasi AL. 2000. The large-scale organization of metabolic networks. Nature. 407:651-654.

Jordan IK, Wolf YI, Koonin EV. 2004. Duplicated genes evolve slower than singletons despite the initial rate increase. BMC Evol Biol. 4:22.

Jovelin R, Dunham JP, Sung FS, Phillips PC. 2009. High nucleotide divergence in developmental regulatory genes contrasts with the structural elements of olfactory pathways in caenorhabditis. Genetics. 181:1387-1397.

Jui HY, Suzuki Y, Accili D, Taylor SI. 1994. Expression of a cDNA encoding the human insulin receptor-related receptor. J Biol Chem. 269:22446-22452.

Kacser H, Burns JA. 1973. The control of flux. Symp Soc Exp Biol. 27:65-104.

Kim PM, Lu LJ, Xia Y, Gerstein MB. 2006. Relating three-dimensional structures to protein networks provides evolutionary insights. Science. 314:1938-1941.

Kimura M. 1983. The neutral theory of molecular evolution. Cambridge University Press, Cambridge.

Kimura M. 1968. Evolutionary rate at the molecular level. Nature. 217:624-626.

King-Jones K, Thummel CS. 2005. Nuclear receptors--a perspective from Drosophila. Nat Rev Genet. 6:311-323.

Koonin EV, Wolf YI. 2006. Evolutionary systems biology: links between gene evolution and function. Curr Opin Biotechnol. 17:481-487.

Krylov DM, Wolf YI, Rogozin IB, Koonin EV. 2003. Gene loss, protein sequence divergence, gene dispensability, expression level, and interactivity are correlated in eukaryotic evolution. Genome Res. 13:2229-2235.

Kuma K, Iwabe N, Miyata T. 1995. Functional constraints against variations on molecules from the tissue level: slowly evolving brain-specific genes

demonstrated by protein kinase and immunoglobulin supergene families. Mol Biol Evol. 12:123-130.

LaPorte DC, Walsh K, Koshland DE, Jr. 1984. The branch point effect. Ultrasensitivity and subsensitivity to metabolic control. J Biol Chem. 259:14068-14075.

Larracuente AM et al. 2008. Evolution of protein-coding genes in Drosophila. Trends Genet. 24:114-123.

Leevers SJ, Weinkove D, MacDougall LK, Hafen E, Waterfield MD. 1996. The Drosophila phosphoinositide 3-kinase Dp110 promotes cell growth. EMBO J. 15:6584-6594.

Lemos B, Bettencourt BR, Meiklejohn CD, Hartl DL. 2005. Evolution of proteins and gene expression levels are coupled in Drosophila and are independently associated with mRNA abundance, protein length, and number of protein-protein interactions. Mol Biol Evol. 22:1345-1354.

Lemos B, Meiklejohn CD, Hartl DL. 2004. Regulatory evolution across the protein interaction network. Nat Genet. 36:1059-1060.

LeRoith D, Olefsky JM, Taylor SI. 2000. Diabetes mellitus: a fundamental and clinical text. Lippincott Williams & Wilkins, Philadelphia.

LeRoith D, Taylor SI, Olefsky JM. 2004. Diabetes mellitus: a fundamental and clinical text. Lippincott Williams & Wilkins, Philadelphia.

Li WH, Wu CI, Luo CC. 1985. A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. Mol Biol Evol. 2:150-174.

Livingstone K, Anderson S. 2009. Patterns of variation in the evolution of carotenoid biosynthetic pathway enzymes of higher plants. J Hered. 100:754-761.

Lu Y, Rausher MD. 2003. Evolutionary rate variation in anthocyanin pathway genes. Mol Biol Evol. 20:1844-1853.

Lynch M, Conery JS. 2003. The origins of genome complexity. Science. 302:1401-1404.

Lynch M, Conery JS. 2000. The evolutionary fate and consequences of duplicate genes. Science. 290:1151-1155.

Marais G, Domazet-Loso T, Tautz D, Charlesworth B. 2004. Correlated evolution of synonymous and nonsynonymous sites in Drosophila. J Mol Evol. 59:771-779.

Marais G, Nouvellet P, Keightley PD, Charlesworth B. 2005. Intron size and exon evolution in Drosophila. Genetics. 170:481-485.

Markow TA, O'Grady PM. 2007. Drosophila biology in the genomic age. Genetics. 177:1269-1276.

Markow TA, O'Grady PM. 2005. Drosophila: a guide to species identification and use. Academic Press, Amsterdam ; Oxford.

Meneses P, De Los Angeles Ortiz M. 1975. A protein extract from Drosophila melanogaster with insulin-like activity. Comp Biochem Physiol A Comp Physiol. 51:483-485.

Mikkelsen TS et al. 2007. Genome of the marsupial Monodelphis domestica reveals innovation in non-coding sequences. Nature. 447:167-177.

Miron M, Lasko P, Sonenberg N. 2003. Signaling from Akt to FRAP/TOR targets both 4E-BP and S6K in Drosophila melanogaster. Mol Cell Biol. 23:9117-9126.

Mouse Genome Sequencing Consortium. 2002. Initial sequencing and comparative analysis of the mouse genome. Nature. 420:520-562.

Nakae J, Kido Y, Accili D. 2001. Distinct and overlapping functions of insulin and IGF-I receptors. Endocr Rev. 22:818-835.

Ohta T. 1973. Slightly deleterious mutant substitutions in evolution. Nature. 246:96-98.

Oldham S, Bohni R, Stocker H, Brogiolo W, Hafen E. 2000. Genetic control of size in Drosophila. Philos Trans R Soc Lond B Biol Sci. 355:945-952.

Oldham S, Hafen E. 2003. Insulin/IGF and target of rapamycin signaling: a TOR de force in growth control. Trends Cell Biol. 13:79-85.

Olsen KM, Womack A, Garrett AR, Suddith JI, Purugganan MD. 2002. Contrasting evolutionary forces in the Arabidopsis thaliana floral developmental pathway. Genetics. 160:1641-1650.

Orian A et al. 2003. Genomic binding by the Drosophila Myc, Max, Mad/Mnt transcription factor network. Genes Dev. 17:1101-1114.

Orme MH, Alrubaie S, Bradley GL, Walker CD, Leevers SJ. 2006. Input from Ras is required for maximal PI(3)K signalling in Drosophila. Nat Cell Biol. 8:1298-1302.

Pál C, Papp B, Hurst LD. 2001. Highly expressed genes in yeast evolve slowly. Genetics. 158:927-931.

Pál C, Papp B, Lercher MJ. 2006. An integrated view of protein evolution. Nat Rev Genet. 7:337-348.

Papadopoulou D, Bianchi MW, Bourouis M. 2004. Functional studies of shaggy/glycogen synthase kinase 3 phosphorylation sites in Drosophila melanogaster. Mol Cell Biol. 24:4909-4919.

Parker J. 1989. Errors and alternatives in reading the universal genetic code. Microbiol Rev. 53:273-298.

Parmley JL, Urrutia AO, Potrzebowski L, Kaessmann H, Hurst LD. 2007. Splicing and the evolution of proteins in mammals. PLoS Biol. 5:e14.

Petit N, Barbadilla A. 2009. The efficiency of purifying selection in Mammals vs. Drosophila for metabolic genes. J Evol Biol. 22:2118-2124.

Ponting CP. 2008. The functional repertoires of metazoan genomes. Nat Rev Genet. 9:689-698.

Potter CJ, Pedraza LG, Xu T. 2002. Akt regulates growth by directly phosphorylating Tsc2. Nat Cell Biol. 4:658-665.

Puig O, Marr MT, Ruhf ML, Tjian R. 2003. Control of cell number by Drosophila FOXO: downstream and feedback regulation of the insulin receptor pathway. Genes Dev. 17:2006-2020.

Ramsay H, Rieseberg LH, Ritland K. 2009. The correlation of evolutionary rate with pathway position in plant terpenoid biosynthesis. Mol Biol Evol. 26:1045-1053.

Rausher MD, Lu Y, Meyer K. 2008. Variation in constraint versus positive selection as an explanation for evolutionary rate variation among anthocyanin genes. J Mol Evol. 67:137-144.

Rausher MD, Miller RE, Tiffin P. 1999. Patterns of evolutionary rate variation among genes of the anthocyanin biosynthetic pathway. Mol Biol Evol. 16:266-274.

Richards S et al. 2005. Comparative genome sequencing of Drosophila pseudoobscura: chromosomal, gene, and cis-element evolution. Genome Res. 15:1-18.

Riley RM, Jin W, Gibson G. 2003. Contrasting selection pressures on components of the Ras-mediated signal transduction pathway in Drosophila. Mol Ecol. 12:1315-1323.

Rocha EP. 2006. The quest for the universals of protein evolution. Trends Genet. 22:412-416.

Rocha EP, Danchin A. 2004. An analysis of determinants of amino acids substitution rates in bacterial proteins. Mol Biol Evol. 21:108-116.

Ruan Y, Chen C, Cao Y, Garofalo RS. 1995. The Drosophila insulin receptor contains a novel carboxyl-terminal extension likely to play an important role in signal transduction. J Biol Chem. 270:4236-4243.

Russo CA, Takezaki N, Nei M. 1995. Molecular phylogeny and divergence times of drosophilid species. Mol Biol Evol. 12:391-404.

Sanger F, Thompson EO. 1953a. The amino-acid sequence in the glycyl chain of insulin. II. The investigation of peptides from enzymic hydrolysates. Biochem J. 53:366-374.

Sanger F, Thompson EO. 1953b. The amino-acid sequence in the glycyl chain of insulin. I. The identification of lower peptides from partial hydrolysates. Biochem J. 53:353-366.

Sanger F, Tuppy H. 1951a. The amino-acid sequence in the phenylalanyl chain of insulin. 2. The investigation of peptides from enzymic hydrolysates. Biochem J. 49:481-490.

Sanger F, Tuppy H. 1951b. The amino-acid sequence in the phenylalanyl chain of insulin. I. The identification of lower peptides from partial hydrolysates. Biochem J. 49:463-481.

Genome 10K Community of Scientists. 2009. Genome 10K: a proposal to obtain whole-genome sequence for 10,000 vertebrate species. J Hered. 100:659-674.

Sharkey TD et al. 2005. Evolution of the isoprene biosynthetic pathway in kudzu. Plant Physiol. 137:700-712.

Sharp PM. 1991. Determinants of DNA sequence divergence between Escherichia coli and Salmonella typhimurium: codon usage, map position, and concerted evolution. J Mol Evol. 33:23-33.

Sharp PM, Li WH. 1989. On the rate of DNA sequence evolution in Drosophila. J Mol Evol. 28:398-402.

Singh ND, Larracuente AM, Sackton TB, Clark AG. 2009. Comparative genomics on the Drosophila phylogenetic tree. Annu. Rev. Ecol. Syst. 40:459-480.

Skorokhod A et al. 1999. Origin of insulin receptor-like tyrosine kinases in marine sponges. Biol Bull. 197:198-206.

Stephanopoulos G, Vallino JJ. 1991. Network rigidity and metabolic engineering in metabolite overproduction. Science. 252:1675-1681.

Subramanian S, Kumar S. 2004. Gene expression intensity shapes evolutionary rates of the proteins encoded by the vertebrate genome. Genetics. 168:373-381.

Taguchi A, White MF. 2008. Insulin-like signaling, nutrient homeostasis, and life span. Annu Rev Physiol. 70:191-212.

Tamura K, Subramanian S, Kumar S. 2004. Temporal patterns of fruit fly (Drosophila) evolution revealed by mutation clocks. Mol Biol Evol. 21:36-44.

Teleman AA. 2010. Molecular mechanisms of metabolic regulation by insulin in Drosophila. Biochem J. 425:13-26.

Teleman AA, Chen YW, Cohen SM. 2005. Drosophila Melted modulates FOXO and TOR activity. Dev Cell. 9:271-281.

Teleman AA, Hietakangas V, Sayadian AC, Cohen SM. 2008. Nutritional control of protein biosynthetic capacity by insulin via Myc in Drosophila. Cell Metab. 7:21-32.

The Bovine Genome Sequencing and Analysis Consortium. 2009. The genome sequence of taurine cattle: a window to ruminant biology and evolution. Science. 324:522-528.

The Honeybee Genome Sequencing Consortium. 2006. Insights into social insects from the genome of the honeybee Apis mellifera. Nature. 443:931-949.

Tuller T, Kupiec M, Ruppin E. 2008. Evolutionary rate and gene expression across different brain regions. Genome Biol. 9:R142.

Van Der Heide LP, Hoekman MF, Smidt MP. 2004. The ins and outs of FoxO shuttling: mechanisms of FoxO translocation and transcriptional regulation. Biochem J. 380:297-309.

van Raalte DH, Ouwens DM, Diamant M. 2009. Novel insights into glucocorticoid-mediated diabetogenic effects: towards expansion of therapeutic options? Eur J Clin Invest. 39:81-93.

Venter JC et al. 2001. The sequence of the human genome. Science. 291:1304-1351.

Vinciguerra M, Foti M. 2006. PTEN and SHIP2 phosphoinositide phosphatases as negative regulators of insulin signalling. Arch Physiol Biochem. 112:89-104.

Wagner A, Fell DA. 2001. The small world inside large metabolic networks. Proc Biol Sci. 268:1803-1810.

Warren WC et al. 2008. Genome analysis of the platypus reveals unique signatures of evolution. Nature. 453:175-183.

Watt WB, Dean AM. 2000. Molecular-functional studies of adaptive genetic variation in prokaryotes and eukaryotes. Annu Rev Genet. 34:593-622.

Weinkove D, Leevers SJ, MacDougall LK, Waterfield MD. 1997. p60 is an adaptor for the Drosophila phosphoinositide 3-kinase, Dp110. J Biol Chem. 272:14606-14610.

Whelan S, Goldman N. 1999. Distributions of Statistics Used for the Comparison of Models of Sequence Evolution in Phylogenetics. Mol Biol Evol. 16:1292-1299.

Wong WS, Yang Z, Goldman N, Nielsen R. 2004. Accuracy and power of statistical methods for detecting adaptive evolution in protein coding sequences and for identifying positively selected sites. Genetics. 168:1041-1051.

Yamaguchi T, Fernandez R, Roth RA. 1995. Comparison of the signaling abilities of the Drosophila and human insulin receptors in mammalian cells. Biochemistry. 34:4962-4968.

Yang J, Gu Z, Li WH. 2003. Rate of protein evolution versus fitness effect of gene deletion. Mol Biol Evol. 20:772-774.

Yang Q, Inoki K, Kim E, Guan KL. 2006. TSC1/TSC2 and Rheb have different effects on TORC1 and TORC2 activity. Proc Natl Acad Sci U S A. 103:6811-6816.

Yang YH, Zhang FM, Ge S. 2009. Evolutionary rate patterns of the Gibberellin pathway genes. BMC Evol Biol. 9:206.

Yang Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci. 13:555-556.

Yang Z, Nielsen R, Goldman N, Pedersen AM. 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics. 155:431-449.

Yenush L et al. 1996. The Drosophila insulin receptor activates multiple signaling pathways but requires insulin receptor substrate proteins for DNA synthesis. Mol Cell Biol. 16:2509-2517.

Yu H et al. 2008. High-quality binary protein interaction map of the yeast interactome network. Science. 322:104-110.

Zhang Z, Parsch J. 2005. Positive correlation between evolutionary rate and recombination rate in Drosophila genes with male-biased expression. Mol Biol Evol. 22:1945-1947.

Zielinski R et al. 2009. The crosstalk between EGF, IGF, and Insulin cell signaling pathways--computational and experimental analysis. BMC Syst Biol. 3:88.

7. Annexos

Characterization of the insulin/TOR pathway genes in the *Pediculus humanus humanus* genome

David Alvarez-Ponce*, Sara Guirao*, Montserrat Aguadé and Julio Rozas

Departament de Genètica, Facultat de Biologia, Universitat de Barcelona, Barcelona 08028, Spain.

Report to the Body Louse Genome Consortium (BLGC)

Work done by the Molecular Evolutionary Genetics group relative to the insulin/TOR signaling pathway genes in the *Pediculus humanus humanus* genome.

*These authors contributed equally to this work.

The insulin/TOR signal transduction pathway plays a central role in multiple and critical biological processes, including organismal growth, anabolic metabolism, cell survival, fertility, and lifespan determination (Goberdhan and Wilson 2003; Oldham and Hafen 2003). This pathway has been well characterized in multiple organisms, including *Drosophila melanogaster.* Both the structure of the pathway and the molecular function of its components are well conserved across Metazoans.

Materials and Methods

The protein coding sequences (CDSs) of the insulin/TOR pathway genes of the *Drosophila melanogaster* genome (release 5.1) were retrieved from the FlyBase database (Crosby et al. 2007). Homologous genes in *Pediculus humanus humanus* were identified by a two-round similarity search (Alvarez-Ponce, Aguadé and Rozas 2009). For each *D. melanogaster* protein, we first performed a TBLASTN search against the body louse genome. Secondly, each hit (*E*-value $\leq 10^{-5}$) was *in silico* translated and used as query for searching the *D. melanogaster* genome. If the best hit of the second round was the original *D. melanogaster* gene, the sequence was considered an orthologous sequence. We also searched for the presence of predicted transcripts and EST *Pediculus* information (search conducted by TBLASTN using the *D. melanogaster* insulin/TOR signaling proteins as a query).

In the case of the *dilp* (*Drosophila insulin-like peptide*) genes that are highly divergent, our efforts to identify these genes by BLAST searches were unsuccessful. We then searched the *P. h. humanus* proteome for the characteristic pattern of the encoded protein (a specific number of cysteines separated by a specific number of residues in both the A and B peptides), which has been observed in vertebrates as well as in most invertebrate species (Claeys et al. 2002; Smit et al. 1998).

Each candidate gene was evaluated manually. We checked for the presence of start and stop codons; those candidate genes with initially partial or erroneous sequence information were reannotated. The intron-exon structure was established: i) by using information from the multiple sequence alignments of known insect insulin signaling genes as well as EST information when available, and ii) using the Splice Site Prediction Server (Neural Network server; http://www. fruitfly.org/seq_tools/splice.html).

Results and Discussion

In order to analyze the body louse insulin/TOR pathway genes, the orthologs of the *Drosophila melanogaster* insulin/TOR genes in the *Pediculus humanus humanus* genome were identified and manually evaluated. The body louse has orthologs for all insulin/TOR pathway *D. melanogaster* genes (table 1) and, therefore, the body louse genome would encode a complete and functional insulin/TOR pathway. However, the number of genes was lower in the body louse than in *D. melanogaster*, with a reduction in the number of those genes with multiple copies in *D. melanogaster* (table 1). Indeed, in *D. melanogaster* fourteen insulin/TOR pathway genes are single-copy, whereas the rest belong to two paralogous groups: seven genes encoding the *Drosophila insulin-like peptides* (*dilp1–7*), and another seven genes encoding the elongation initiation factors 4E (*eIF-4E*, *eIF4E3–7* and *4EHP*). Remarkably, the *P. h. humanus* genome has a single insulin-like peptide (*ilp*) gene. Given that there is some evidence for differential expression of *ilp* genes under different dietary conditions in insects (Wheeler et al. 2006; Arsic and Guerin 2008), the presence of a single *ilp* gene in the body louse genome might reflect its restricted and homogeneous diet. Also, the *P. h. humanus* genome contains three eIF4E-encoding genes, one of each of the three *eIF4E* gene classes described in Joshi et al. (2005), which contrasts with the absence of class III genes in Diptera.

Additional Information

eIF4E family members

eIF4E proteins control the recruitment of the majority of eukaryotic mRNAs to the ribosome through binding to their 5'-m^7Gppp cap-structure and the elF4G scaffolding protein. Although eIF4E proteins seem to be present in all eukaryotes, the number of genes encoding these proteins substantially varies among species. Non-protist eIF4E family members have been classified into three classes differing in structure, function and phylogenetic distribution across eukaryotes (Joshi et al. 2005): class I, with Trp at residues 43 and 56 of the human eIF4E-1 and present presumably in all eukaryotes; class II, eIF4Es with Trp43Tyr/Phe/Leu and Trp56Tyr/Phe substitutions and present in Metazoans, Viridiplantae and Fungi; and class III, with Trp at residue 43 and Trp56Cys/Tyr substitutions, and well represented in chordates and sporadically represented among invertebrates. Class I members have all the functions associated to the eIF4E proteins

whereas, in Metazoans, the other two classes are thought to have only a subset of these functions (Joshi et al. 2004; Hernandez et al. 2005).

References

Alvarez-Ponce D, Aguadé M, Rozas J. 2009. Network-level molecular evolutionary analysis of the insulin/TOR signal transduction pathway across 12 Drosophila genomes. Genome Res 19(2): 234-242.

Arsic D, Guerin PM. 2008. Nutrient content of diet affects the signaling activity of the insulin/target of rapamycin/p70 S6 kinase pathway in the African malaria mosquito Anopheles gambiae. J Insect Physiol 54(8): 1226-1235.

Claeys I, Simonet G, Poels J, Van Loy T, Vercammen L, De Loof A, Vanden Broeck J. 2002. Insulin-related peptides and their conserved signal transduction pathway. Peptides 23: 807-816.

Crosby MA, Goodman JL, Strelets VB, Zhang P, Gelbart WM. 2007. FlyBase: genomes by the dozen. Nucleic Acids Res 35: D486-491.

Goberdhan DC, Wilson C. 2003. The functions of insulin signaling: size isn't everything, even in Drosophila. Differentiation 71: 375-397.

Hernandez G, Altmann M, Sierra JM, Urlaub H, Diez del Corral R, Schwartz P, Rivera-Pomar R. 2005. Functional analysis of seven genes encoding eight translation initiation factor 4E (eIF4E) isoforms in Drosophila. Mech Dev 122:529-543.

Joshi, B, Cameron A, Jagus R. 2004. Characterization of mammalian eIF4E-family members. Eur J Biochem 271: 2189-2203.

Joshi B, Lee K, Maeder DL, Jagus R. 2005. Phylogenetic analysis of eIF4E-family members. BMC Evol Biol 5: 48.

Oldham S, Hafen E. 2003. Insulin/IGF and target of rapamycin signaling: a TOR de force in growth control. Trends Cell Biol 13: 79-85.

Smit AB, van Kesteren RE, Li KW, Van Minnen J, Spijker S, Van Heerikhuizen H, Geraerts WP. 1998. Towards understanding the role of insulin in the brain: lessons from insulin-related signaling systems in the invertebrate brain. Prog Neurobiol 54: 35-54.

Wheeler DE, Buck N, Evans JD. 2006. Expression of insulin pathway genes during the period of caste determination in the honey bee, Apis mellifera. Insect Mol Biol 15(5): 597-602.

Table 1. Insulin/Tor pathway genes detected in the *Pediculus humanus humanus* genome as compared to those present in the *Drosophila melanogaster* genome

		D. melanogaster	***P. h. humanus***			
Gene	**Protein**	**Number of genes**	**Number of genes**	**Type of evidence**		
				BLAST search	**ESTs (*P. h. humanus*)**	**ESTs (*P. h. capitis*)**
Akt1	PKB	1	1	Yes	No	Yes
chico	Chico	1	1	Yes	No	No
dilps	Dilps	7	1[a]	No	No	No
***eIF-4E* (class I)**	eIF4E-1–7	6	1	Yes	No	No
***eIF-4E* (class II)**	eIF4E-8	1	1	Yes	No	No
***eIF-4E* (class III)**	-	0	1	Yes	No	No
gig	Tsc2	1	1	Yes	No	Yes
InR	InR	1	1	Yes	No	No
Pi3K21B	p60	1	1	Yes	No	No
Pi3K92E	p110	1	1	Yes	No	No
Pk61C	PDK1	1	1	Yes	No	No
Pten	PTEN	1	1	Yes	No	Yes
Rheb	Rheb	1	1	Yes	No	No
RpS6	RpS6	1	1	Yes	Yes	No
S6k	S6k	1	1	Yes	No	No
Thor	d4E-BP	1	1	Yes	No	No
Tor	TOR	1	1	Yes	No	No
Tsc1	Tsc1	1	1	Yes	No	No
Total		**28**	**18**			

[a] The *P. h. humanus insulin-like peptide* gene was identified searching in the body louse proteome with the peptide characteristic amino acid pattern (Smit et al. 1998; Claeys et al. 2002).

UNIVERSITAT DE BARCELONA

Departament de Genètica
Facultat de Biologia
Diagonal 645
08028 Barcelona
Spain

Julio Rozas
☎ +34 93 402 1495
Fax: +34 93 403 4420
E-mail: jrozas@ub.edu

Informe del director de la tesi especificant la participació feta pel doctorand en cada article, i si algun dels coautors d'algun dels treballs presentats en la tesi doctoral ha utilitzat, implícitament o explícitament aquests treballs per a l'elaboració d'una tesi doctoral

---

Els Drs. **Montserrat Aguadé Porres** i **Julio Rozas Liras,** directors de la Tesi Doctoral elaborada pel Sr. **David Álvarez Ponce**, amb el títol "**Genòmica evolutiva de la via de transducció de senyal de la insulina/TOR a insectes i vertebrats",**

INFORMEN

Que la tesi doctoral està elaborada com a compendi de 2 publicacions amb dades originals (publicacions 1-2 en el cos central de la tesi), i una tercera a l'apèndix:

1. Alvarez-Ponce, D., M. Aguadé & J. Rozas. 2009. Network-level molecular evolutionary analysis of the insulin/TOR signal transduction pathway across 12 *Drosophila* genomes. *Genome Res.* **19:** 234-242. Factor d'impacte: **10.176**. Ocupa la posició **7** (sobre **138**) dins la categoria de Genetics and Heredity.

2. Alvarez-Ponce, D., M. Aguadé & J. Rozas. 2010. Comparative genomics of the vertebrate insulin/TOR signal transduction pathway genes: A network-level analysis of selective pressures along the pathway. (En preparació). Es vol enviar a la revista *Genome Biology and Evolution*.

3. Alvarez-Ponce, D., S. Guirao, M. Aguadé & J. Rozas. "Characterization of the insulin/TOR pathway genes in the *Pediculus humanus humanus* genome". Aquest informe s'ha enviat al Body Louse Genome Consortium, i és la base de la contribució dels seus 4 autors a l'article de Kirkness *et al.* (2010) que es publicarà als Proc. Natl. Acad. Sci. USA (en premsa). Factor d'impacte: **9.380**. Ocupa la posició **3** (sobre **42**) dins la categoria de Multidisciplinary Sciences.

A les publicacions 1 i 2 el doctorand va realitzar la feina computacional i d'anàlisi de dades, i va redactar el primer esborrany dels manuscrits. A la publicació 3, on participen diversos grups de recerca, va dur a terme amb S. Guirao (estudiant de doctorat de M. Aguadé) l'anàlisi dels gens de la via de la insulina/TOR. Tant S. Guirao com D. Álvarez han inclòs aquest informe a l'apèndix de les seves respectives tesis, i per tant no forma part del cos central de les mateixes.

Dra. Montserrat Aguadé Porres Dr. Julio Rozas Liras

Aquesta tesi doctoral ha estat finançada pels projectes BFU2004-02253, BFU2007-62927 i BFU2007-63228 del Ministerio de Educación y Ciencia, i pels projectes 2005SGR-00166 i 2009SGR-1287 de la Comissió Interdepartamental de Recerca i Innovació Tecnològica. Durant el període de formació pre-doctoral, DAVID ÁLVAREZ PONCE ha gaudit d'una Beca per a la Formació de Personal Investigador (FI) de la Generalitat de Catalunya (3 mesos), una Beca de Formación del Profesorado Universitario (FPU, ref. AP2005-0012), del Ministerio de Educación y Ciencia (45 mesos), i de dues beques de col·laboració amb projectes d'investigació del grup de Genètica Molecular Evolutiva del Departament de Genètica de la Universitat de Barcelona (7 mesos en total).