Genómica evolutiva de la regulación transcripcional en las principales familias multigénicas del sistema quimiosensorial de *Drosophila*

Pablo Librado Sanz

Universitat de Barcelona

Facultat de Biologia
Departament de Genètica

Genómica evolutiva de la regulación transcripcional en las principales familias multigénicas del sistema quimiosensorial de *Drosophila*

Pablo Librado Sanz

Barcelona, Febrero de 2014

Genómica evolutiva de la regulación transcripcional en las principales familias multigénicas del sistema quimiosensorial de *Drosophila*

Memoria presentada por Pablo Librado Sanz
para optar al Grado de Doctor
por la Universitat de Barcelona.

Departament de Genètica
Universitat de Barcelona

El director y tutor de la tesis
Dr. Julio Rozas Liras
Catedràtic de Genètica
Universitat de Barcelona

El autor de la tesis
Pablo Librado Sanz

Barcelona, Febrero de 2014

A mis padres

Índice

1 Introducción 1

1.1 Evolución biológica mediante selección natural 1

1.1.1 La huella molecular de la selección natural 1

1.2 Regulación de la transcripción 3

1.2.1 Elementos *cis*-reguladores de la transcripción 4

1.2.2 Dominios de la cromatina 5

1.3 Duplicación génica y familias multigénicas 6

1.4 El sistema quimiosensorial 8

1.4.1 El sistema quimiosensorial periférico en *Drosophila* 9

1.4.2 Las familias multigénicas del sistema quimiosensorial en *Drosophila* 10

1.5 Evolución de las familias multigénicas del sistema quimiosensorial 12

1.5.1 Regiones codificadoras 12

1.5.2 Regiones reguladoras de la transcripción 14

1.5.3 Ganancia y pérdida de genes quimiosensoriales 15

1.5.4 Distribución genómica de las OBPs 16

1.6 Especies y poblaciones analizadas 19

1.6.1 Las 12 especies de *Drosophila* 19

1.6.2 *Drosophila Genetic Reference Panel* 20

2 Objetivos 21

3 Capítulos 23

3.1 DnaSP v5: a software for comprehensive analysis of DNA polymorphism data 23

3.2 BadiRate: estimating family turnover rates by likelihood-based methods 27

3.3 PopDrowser: the Population Drosophila Browser 47

3.4 Uncovering the functional constraints underlying the genomic organization of the odorant-binding protein genes 51

3.5 Positive selection drives the evolution of the transcriptional regulatory upstream regions of the major chemosensory gene families 73

4 Discusión 121
4.1 Implementación de nuevos métodos analíticos 121
4.1.1 DnaSP v5 121
4.1.2 BadiRate 123
4.1.3 popDrowser 124
4.2 Evolución de la regulación transcripcional de los genes quimiosensoriales 125
4.2.1 Distribución física de los genes que codifican OBPs 125
4.2.2 Evolución de las regiones *upstream* 128

5 Conclusiones 133

Bibliografía

Bibliografía 137

Anexo

A El proyecto DGRP 149

B Mycobacterial phylogenomics 157

C Informe del director 209

D Financiación 213

1
Introducción

1.1 Evolución biológica mediante selección natural

El concepto de evolución biológica fue acuñado por algunos filósofos griegos [1, 2], e incluso por el mismo Charles Bonnet en el siglo XVIII [3]. No obstante, hasta que en 1859 no se publicaron las obras de Charles Darwin y Alfred Russel Wallace (donde se recogen múltiples evidencias empíricas) [4], no hubo un cambio de paradigma en el pensamiento científico y filosófico: la biodiversidad existente no es la obra de un creador divino, sino el resultado de un proceso natural de cambios graduales heredados de un antepasado común (evolución biológica; Figura 1.1).

¿Cuáles son los mecanismos responsables de la evolución biológica? Los principales motores de la evolución son la mutación, la recombinación, la deriva genética y la selección natural. Aunque todos ellos interactúan simultáneamente, sólo la selección natural puede explicar la adaptación de las especies al medio, a través de la reproducción diferencial de los individuos. Efectivamente, los individuos portadores de mutaciones deletéreas tendrán menos oportunidad de reproducirse, de tal forma que la frecuencia de sus mutaciones disminuirá en la población (selección negativa o purificadora). Al contrario, las mutaciones beneficiosas incrementarán su frecuencia (selección positiva o *darwiniana*), pudiendo *fijarse* en todos los individuos de la especie.

1.1.1 La huella molecular de la selección natural

La determinación del impacto de la selección natural es fundamental en biología evolutiva. Además de poseer profundas implicaciones conceptuales, su

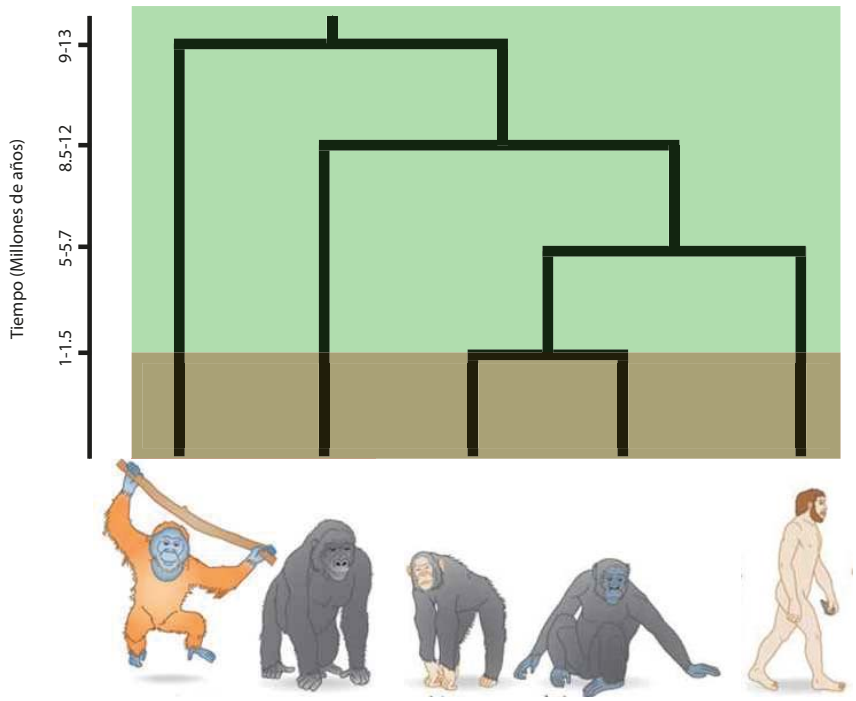


Figura 1.1: Cambio de paradigma en el pensamiento biológico. Los creacionistas (visión 'limitada' por el recuadro marrón) se ciñen -paradójicamente- a la realidad directamente observable: 'existen diferentes especies y por tanto deben haber existido siempre (son inmutables)'. Los evolucionistas, sin embargo, entienden que las especies descienden de un ancestro común, a través de cambios graduales. Adaptado de [5].

detección puede ayudarnos a comprender la función de los elementos genéticos (ej. genes), así como también revelar que mutaciones explican los procesos adaptativos. Uno de los ejemplos más mediáticos es la adaptación de las poblaciones humanas a la vida en altitudes superiores a los 4000 metros, donde la concentración de oxígeno es un 40 % menor que en el nivel del mar. Recientes estudios han constatado que un gran porcentaje de los individuos de la meseta tibetana portan mutaciones relacionadas con el metabolismo del oxígeno, como una mutación cercana al gen *EPAS1* (coloquialmente conocido como 'gen de los súper atletas').

¿Cómo se puede discernir el impacto de la selección natural de los otros mecanismos responsables de la evolución biológica? La teoría de la coalescencia provee el marco matemático-estadístico necesario para contrastar ésta y otras hipótesis. Desafortunadamente, la modelización de muchos escenarios evolutivos no dispone de solución analítica, haciendo imprescindible la búsqueda de soluciones numéricas mediante simulaciones por ordenador (computacio-

nalmente muy costosas).
Una aproximación igualmente popular es la comparación de dos clases de posiciones: una focal (potencial diana de la selección natural) y una neutra (donde las mutaciones no tienen consecuencias fenotípicas y, por tanto, son 'invisibles' para la selección natural) [6-8]. Así, en las regiones codificadoras de proteína, las mutaciones se clasifican en no sinónimas (clase focal) o sinónimas (clase neutra) dependiendo de si comportan un cambio de aminoácido o no. De hecho, el ratio entre la divergencia no sinónima (d_N) y sinónima (d_S) es uno de los estadísticos más robustos para inferir el impacto de la selección natural ($\omega = d_N/d_S$); valores de ω que se desvían significativamente de uno (lo esperado bajo un modelo neutro) son interpretados como el efecto de la selección negativa ($\omega < 1$) o positiva ($\omega > 1$) [9, 10].
La mayoría de procesos adaptativos ocurren en episodios puntuales de la historia. Para detectarlos, la clase focal y la neutra se pueden comparar a diferentes tiempos evolutivos: analizando la variación intraespecífica o *polimorfismo* (que permite inferir el impacto reciente de la selección natural), variación interespecífica o *divergencia* (para detectar la huella pretérita de la selección natural) o ambas (*polimorfismo* y *divergencia*, útil para estudiar tiempos evolutivos intermedios). Todas estas metodologías han permitido estimar el impacto de la selección natural en regiones codificadoras [11, 12]. No obstante, los estudios genómicos que abordan la contribución de la regulación transcripcional a los procesos adaptativos todavía son escasos [13].

1.2 Regulación de la transcripción

Las instrucciones genéticas para el correcto desarrollo y funcionamiento del organismo están codificadas en su ADN. Así, los genes contienen la información necesaria para la síntesis de macromoléculas (ej. las proteínas), que son las grandes responsables de acometer las funciones celulares, desde catalizar reacciones metabólicas hasta contraer el músculo.
La comparación del número de genes por especie muestra algunos resultados aparentemente contradictorios. Basándose en la complejidad de organismos previamente secuenciados, algunos miembros de la comunidad científica estimaron que el genoma humano debería contener entre 31000 y 140000 genes. Sin embargo, su secuenciación determinó la existencia de tan sólo 22721 genes [14, 15], un número comparable a los 23300 del erizo de mar [16], pero mucho menor que los 40745 de la planta del arroz [17].
¿Cómo es posible que el genoma de una especie tan 'compleja' como la humana apenas tenga la mitad de genes que el arroz? Entre las posibles explicaciones, destaca la idea de que la complejidad de un organismo no sólo depende del número de macromoléculas codificadas en su genoma, sino también de la

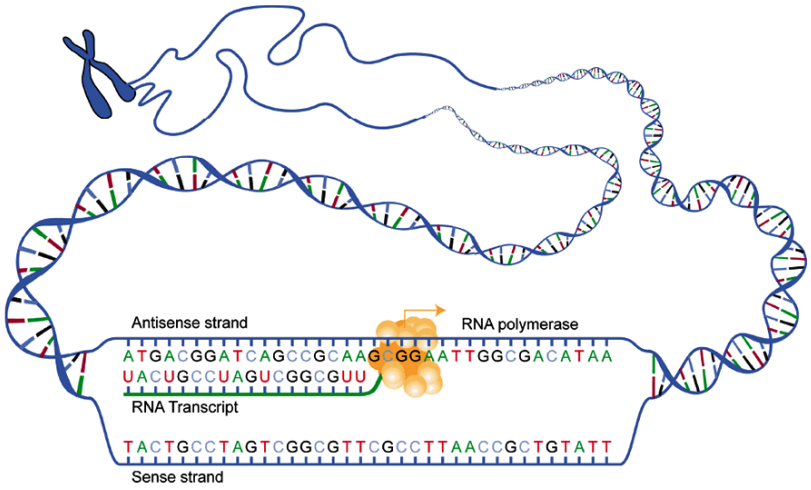


Figura 1.2: La enzima ARN polimerasa 'transcribe' la información codificada en el gen a transcrito mensajero.

cantidad, lugar y momento en el que se sintetizan (expresan). Por tanto, para comprender la biología de los individuos y de las especies es imprescindible examinar los mecanismos que controlan la expresión génica.

La síntesis proteica implica dos procesos moleculares sucesivos: la transcripción del ADN a ARN (Figura 1.2) y la posterior traducción del ARN a proteína. Actualmente, se sabe que la abundancia de proteína se puede controlar a diferentes niveles, mediante mecanismos pre-transcripcionales, post-transcripcionales y post-traduccionales. Tanto por cuestiones metodológicas, como por constituir el primer control de la expresión génica, la comunidad científica se ha centrado en el estudio de los mecanismos pre-transcripcionales, especialmente los que implican elementos *cis*-reguladores (CREs) y dominios de la cromatina.

1.2.1 Elementos *cis*-reguladores de la transcripción

Los CREs son secuencias cortas de ADN (típicamente entre 6 y 15 nucleótidos) donde se unen -de forma más o menos específica- las proteínas reguladoras que inducen o reprimen la actividad transcripcional. Cada CRE puede controlar la transcripción de varios genes y, de forma complementaria, la transcripción de cada gen puede estar regulada por varios CREs. Los CREs que dirigen la transcripción del mismo gen suelen estar físicamente agrupados, formando módulos *cis*-reguladores (CRMs) en su región proximal (promotor) o distal (*enhancers*) [18].

La unión de la ARN polimerasa al promotor normalmente induce un nivel basal de transcripción génica. Entonces, ¿cómo se ajusta la expresión en cada tejido o estadio del desarrollo? Entre otros mecanismos, destaca el que proporcionan los *enhancers*, modulando estos niveles basales en un marco espacio-temporal concreto, sin alterar la expresión en otras condiciones (sin efectos pleitrópicos).
Se ha hipotetizado que -debido a sus reducidos efectos pleitrópicos- las mutaciones en los CRMs pueden ser el principal sustrato de la selección positiva [19]. La lógica que subyace a esta hipótesis es la siguiente: las mutaciones con múltiples efectos pleitrópicos normalmente comprometen la viabilidad del individuo y, por tanto, son purgadas por la selección purificadora. Al contrario, la naturaleza modular de los CRMs (tanto a nivel estructural como funcional) garantiza que las consecuencias fenotípicas de sus mutaciones se circunscriban a un marco espacio-temporal concreto. Uno de los ejemplos más ilustrativos acerca del potencial adaptativo de los CRMs es el conocido como 'persistencia de la lactasa'. A diferencia del resto de mamíferos, algunos humanos expresan dicho enzima más allá de la época de lactancia. La capacidad de digerir la leche durante toda la vida es -probablemente- una adaptación a un cambio sustancial en nuestra actividad socio-económica: la ganadería. Ciertamente, la domesticación de algunas especies animales estabilizó el abastecimiento de nutrientes, lo que tuvo un gran impacto en el desarrollo de las poblaciones humanas. Pues bien, gran parte de este cambio transcripcional se explica por la mutación de un único CRM, situado a unos 14000 nucleótidos del gen que codifica para la enzima lactasa [20]). A pesar de éste y otros ejemplos, la contribución relativa de los CREs al cambio adaptativo sigue siendo una de las principales controversias en biología evolutiva: los críticos abogan por otros mecanismos que también reducen la pleiotropía de las mutaciones, como la duplicación génica [21].

1.2.2 Dominios de la cromatina

¿Cómo es posible que el diámetro del núcleo de las células eucariotas sea aproximadamente un millón de veces menor que la longitud del ADN que alberga? Dentro del núcleo, el ADN no está en una conformación canónica lineal, sino empaquetado, formando la cromatina. La cromatina está constituida por una serie de proteínas, principalmente octámeros de histonas (nucleosomas) alrededor de los cuales se enrolla (empaqueta) el ADN [22].
Las regiones empaquetadas dificultan la accesibilidad de la maquinaria transcripcional a los CREs [23]. Para garantizar que los genes se transcriban en el lugar y el momento adecuado, los eucariotas han adquirido dos mecanismos complementarios. Por una parte, la cromatina puede desempaquetarse localmente [24, 25], facilitando el acceso de la ARN polimerasa y de otras proteínas

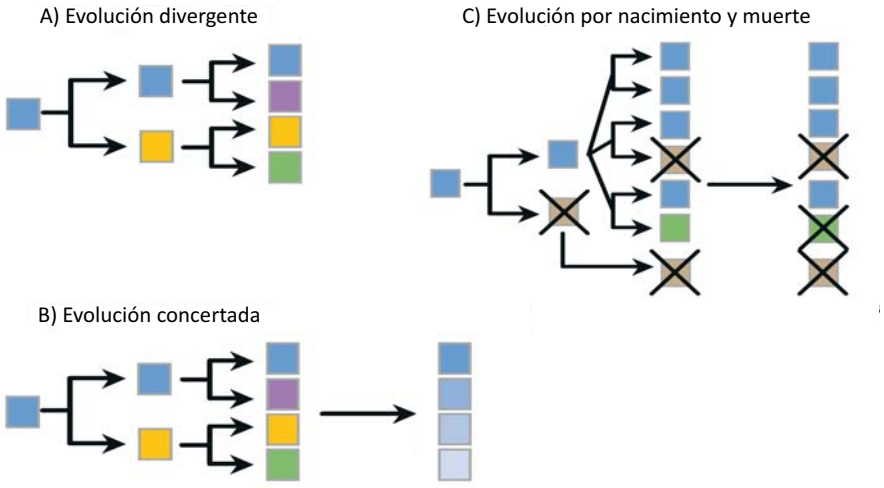


Figura 1.3: Modelos evolutivos de las familias multigénicas. A. Evolución divergente: los miembros de la familia multigénica divergen gradualmente; B. Evolución concertada: las copias génicas homogeneizan su información mediante mecanismos como la conversión génica; C. Nacimiento y muerte de genes: las copias se ganan por duplicación génica, divergen gradualmente y, eventualmente, se pierden. Cada cuadrado representa un gen. Adaptado de [29].

reguladoras. Por otra parte, como la remodelación de la cromatina es energéticamente costosa [26], los genes tienden a estar agrupados en aquellas regiones cromosómicas con un ambiente transcripcional adecuado. Por ejemplo, los centrómeros y telómeros (constitutivamente empaquetados por cuestiones estructurales) presentan una menor densidad génica [27]. Por tanto, el estado de la cromatina no sólo regula la transcripción, sino que también influencia la distribución física de los genes.

1.3 Duplicación génica y familias multigénicas

Una familia multigénica es un conjunto de genes relacionados filogenéticamente (son homólogos), porque derivan de una serie de eventos de duplicación de ADN. Se han propuesto diferentes modelos para la evolución de las familias multigénicas (Figura 1.3): la evolución divergente, la evolución concertada y la evolución por nacimiento y muerte. Aunque no son excluyentes, el último modelo es el que mejor se ajusta a la mayoría de familias multigénicas estudiadas hasta la fecha [28].
En el modelo de nacimiento y muerte, los genes se ganan por duplicación y se pierden por deleción o pseudogenización. Existen diferentes mecanismos moleculares que explican la duplicación de ADN, incluyendo el entrecruzamiento desigual (que produce duplicaciones en tándem) o la retrotransposición.

Inmediatamente después de duplicarse, los genes (parálogos) codifican proteínas con la misma estructura primaria. No obstante, como el proceso mutacional es estocástico, los parálogos divergirán gradualmente con tres posibles destinos: (i) la pseudogenización (una de las copias acumulará mutaciones de pérdida de función); (ii) la neofuncionalización (una de las copias divergirá, hasta adquirir una nueva función); y (iii) la subfuncionalización (las dos copias divergirán, adquiriendo funciones parcialmente complementarias, aunque entre ambas retienen la función ancestral) [30].

Para comprender la contribución relativa de estos mecanismos a la divergencia entre especies, es fundamental estimar la tasa de nacimiento (β) y muerte (δ) de genes. La aproximación más popular es, probablemente, la reconciliación del árbol de genes (AG) con el de especies (AE) [31-33]. Brevemente, la reconciliación intenta explicar las incongruencias del AG con el AE mediante diferentes eventos de duplicación y pérdida de genes. No obstante, es bien sabido que el AG puede ser incongruente con el AE por otros motivos, tanto metodológicos como biológicos. Así, la mayoría de algoritmos de reconciliación obvian -por cuestiones computacionales- la incertidumbre asociada al alineamiento de secuencias o a la propia inferencia filogenética (problemas metodológicos). Además, la reconciliación del AG con el AE puede no representar la verdadera historia de ganancia y pérdida genes si hay: polimorfismo ancestral [34], conversión génica [35], transferencia genética horizontal (HGT) [34] o, simplemente, presiones selectivas heterogéneas entre los diferentes genes [36] (problemas biológicos). En ambas situaciones, el número de eventos de duplicación y, especialmente, de pérdida de genes puede sobrestimarse considerablemente [37].

Actualmente, no existen modelos computacionalmente factibles que integren la incertidumbre metodológica y las posibles incongruencias entre el AG y el AE. Por ello la ganancia y pérdida de genes se estudia mediante modelos de nacimiento y muerte (BD) [38, 39]. Los modelos de BD son estadísticamente complejos. En concreto, son cadenas de *Markov*, donde el número de genes son los posibles estados de la cadena y la transición entre estados es proporcional al número de genes existentes. Efectivamente, la probabilidad que ocurra una duplicación en tándem, una pseudogenización o una deleción incrementa con el número de genes existentes (son mecanismos moleculares dependientes de densidad) (Figura 1.4A).

Debido a que no se puede producir una duplicación si al menos no existe un gen, la probabilidad de transición de cero a un gen es nula (cero es un estado absorbente de la cadena de *Markov*). Que dicha probabilidad sea nula limita la aplicación de los modelos clásicos de BD, por dos motivos. Primero, durante la definición computacional de las familias multigénicas, los miembros muy divergentes quedan separados en subfamilias. Algunas subfamilias son específicas de linaje, con lo que 'aparentemente' se originan *de novo* (de cero a un

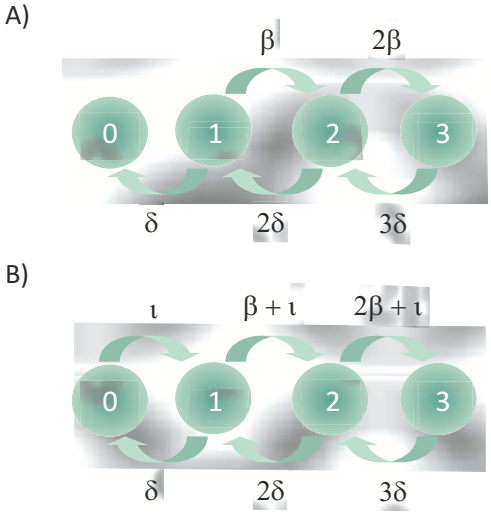


Figura 1.4: A. Los primeros tres estados de la cadena de *Markov* en un modelo BD donde las probabilidades de nacimiento (β) y muerte (δ) son proporcionales al número de genes existentes. En este caso, la transición de 0 a 1 gen es nula; B. Lo equivalente en un modelo BDI, donde la ganancia de elementos tiene un componente dependiente de densidad (β) y otro independiente (ι).

gen). Segundo, existen mecanismos de ganancia de genes que realmente son independientes de densidad; por ejemplo, en procariotas puede haber eventos de HGT, los cuales no dependen del número de genes existentes en el genoma receptor.

Para soslayar estas limitaciones, se desarrollaron los modelos de nacimiento, muerte e innovación/inmigración (BDI) (Figura 1.4B) [39]. En estos modelos, la probabilidad de ganar un gen tiene un componente dependiente de densidad (β) y otro independiente (ι); de hecho, el modelo BD es un caso particular del BDI, donde $\iota = 0$. La inclusión de una tasa independiente de densidad (ι) posibilita analizar familias muy divergentes, casos de HGT e incluso familias de elementos funcionales típicamente cortos, que pueden originarse *de novo*, como los CREs.

1.4 El sistema quimiosensorial

Todos los organismos interaccionan con el ambiente a través de señales, tanto físicas (visuales, acústicas, térmicas, electromagnéticas o táctiles), como químicas [40]. Aunque todas ellas aportan información complementaria, la ingente diversidad de compuestos químicos conlleva un gran potencial para percibir el estado del medio externo. Por ello, procesos tan críticos como la detección

de nutrientes o pareja dependen -en gran medida- del sistema quimiosensorial [41-44]. La quimiopercepción puede dividirse en dos subsistemas fisiológicamente relacionados, pero anatómicamente independientes, el gusto y el olfato. El gusto permite la detección de compuestos solubles, mientras que el olfato de compuestos volátiles. Esta subdivisión funcional está estrechamente vinculada con el medio que habitan las especies. En términos generales, la discriminación de estímulos en fase gaseosa es una adaptación crítica al medio terrestre [45-47], aunque algunos organismos acuáticos también presentan sistema olfativo [48].

Durante la diversificación de los artrópodos, han ocurrido diversos eventos de terrestralización independientes (ej. en insectos, miriápodos, quelicerados y crustáceos), con lo que algunos elementos del sistema olfativo se han tenido que adquirir de forma independiente en cada linaje, a partir de genes ya existentes (cooptación). La ubicuidad y convergencia evolutiva de la percepción olfativa ponen de manifiesto que el sistema quimiosensorial es crítico para la supervivencia de los organismos y, por tanto, un excelente candidato para estudiar los procesos adaptativos a nivel molecular.

1.4.1 El sistema quimiosensorial periférico en *Drosophila*

En *Drosophila*, las primeras etapas de la quimiopercepción ocurren en unas estructuras quitinosas en forma de pelo llamadas sensilios (Figura 1.5A). Los principales componentes de los sensilios son: la linfa (una solución acuosa que baña su interior), las células accesorias y las neuronas receptoras (RNs) [49, 50].

Las características de los sensilios dependen de la naturaleza de los estímulos que procesan. Los sensilios olfativos (incluyen los subtipos tricoideos, coelónicos y basicónicos) presentan una membrana con múltiples poros y están típicamente inervados por dos ORNs (Figura 1.5A). Por el contrario, los gustativos (con los subtipos largos, intermedios y cortos) tienen un único poro en el ápice, y entre dos y cuatro GRNs. Además, también existen diferencias en la localización de los sensilios; mientras los olfativos están restringidos al palpo maxilar y al tercer segmento antenal, los gustativos están ampliamente distribuidos por el cuerpo de la mosca, incluyendo proboscis, patas, alas e incluso placa vaginal [53].

A pesar de las particularidades de los sensilios, hay tres fases que son comunes a todos los procesos de quimiopercepción periférica: (i) la perfusión del estímulo a través de los poros cuticulares (del medio externo hasta la linfa del sensilio), (ii) su difusión a través de la linfa (hasta la membrana de la RN), (iii) y la activación de los quimioreceptores anclados en la membrana de la RN (transmiten la señal a centros neuronales superiores (Figura 1.5B)).

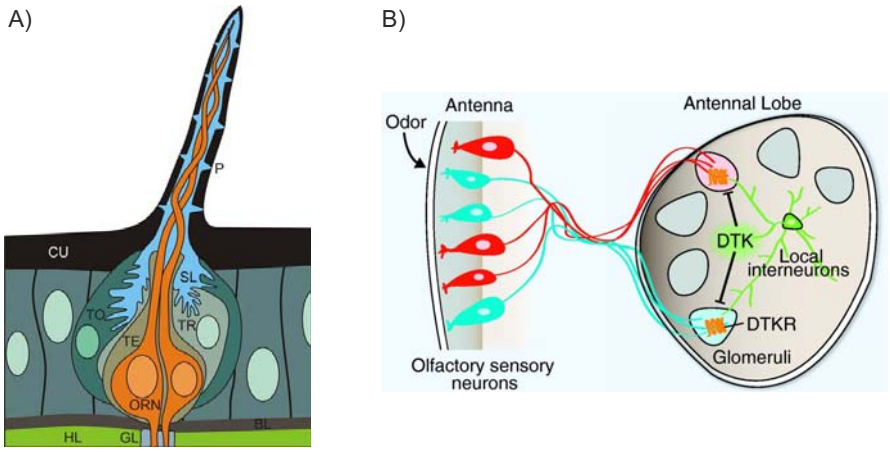


Figura 1.5: A) Sensilio olfativo del tipo tricoideo, con sus respectivos poros (P) en la cutícula (CU), la linfa del sensilio (SL), así como también dos neuronas sensoriales olfativas (ORNs). Fuente: [51]. B) Representación esquemática de la proyección axónica de los ORNs a los glomérulos olfativos del lóbulo antenal. Fuente: [52].

1.4.2 Las familias multigénicas del sistema quimiosensorial en *Drosophila*

Los estudios de genética molecular, conjuntamente con los de comportamiento, han permitido identificar una serie de proteínas involucradas en las primeras etapas de la quimiopercepción. Actualmente, con la creciente disponibilidad de genomas secuenciados y métodos bioinformáticos, se ha constatado que muchas de estas proteínas están codificadas en familias multigénicas [54-56], entre las que destacan las proteínas extracelulares de unión a ligando y los propios quimiorreceptores. No obstante, que los miembros de una familia multigénica tengan un origen común no implica -necesariamente- que todos estén involucrados en la quimiopercepción. Es muy probable que algunos de sus miembros hayan divergido funcionalmente, especializándose en otros procesos biológicos o funciones moleculares relacionadas.

Proteínas de unión a ligando

Las células accesorias de los sensilios quimiosensoriales secretan una gran cantidad de pequeñas proteínas globulares (120-230 aminoácidos, aproximadamente) de unión a ligando, entre las que destacan las *Odorant-Binding Proteins* (OBPs) y las dos familias de *Chemosensory Proteins* (CSPs, CheAs y CheBs) [57-60]. Tanto las CSPs como las OBPs se han clasificado en diferentes subfamilias. Por ejemplo, aparte de las OBPs denominadas clásicas, se han descrito las PBP/GOBP, las minus-C, las plus-C, las diméricas, las ABPI, las ABPII, las

CRLBP y las D7 [61, 62]. Estas subfamilias se han definido en base a criterios filogenéticos, funcionales y estructurales, especialmente por su patrón de cisteínas. La conservación del patrón de cisteínas es crítico para la funcionalidad de la proteína, ya que estabiliza su plegamiento globular, presentando los residuos hidrofílicos en el exterior (facilita su solubilización en la hemolinfa), y los hidrofóbicos alrededor del sitio de unión (posibilita el transporte de estímulos volátiles).

La función de las proteínas de unión a ligando todavía no se conoce completamente. La proteína mejor caracterizada es la Obp76a (lush), que participa en la detección de la feromona 11-cis-vacenil acetato (cVA). Estudios cristalográficos de rayos X han demostrado que cVA se une físicamente a lush, lo que induce un cambio en la conformación de la proteína. Se especula que esa conformación 'activa' de lush es capaz de disparar, directa o indirectamente, la actividad del quimiorreceptor Or67d [63], el cual inerva glomérulo olfativo DA1. No obstante, hay dos evidencias parcialmente contradictorias con esta hipótesis [64]. Primero, mutantes que mimetizan la conformación 'activa' de lush no tienen ningún efecto sobre la actividad del receptor, cuando son expresados *in vivo*. Segundo, altas concentraciones de cVA pueden activar el receptor por sí mismas, sugiriendo que es cVA (y no lush) la que activa Or67d. Así, el modelo actual sugiere que lush no es indispensable para la detección de cVA, pero si importante para incrementar la sensibilidad del sistema olfativo frente a las concentraciones de cVA existentes en la naturaleza.

Parece inverosímil que este modelo funcional sea extrapolable a todos los miembros de estas familias multigénicas. Algunas OBPs y CSPs se expresan en órganos no quimiosensoriales, incluyendo el tarso o las glándulas accesorias de macho [65, 66]. Incluso se ha constatado que algunas CSPs participan en otros procesos biológicos, como en el desarrollo y la regeneración [67]. Además, las CheAs y las CheBs pueden tener un papel adicional a la de solubilizar compuestos hidrofóbicos, ya que se expresan mayoritariamente en los sensilios gustativos, donde participan en la detección de feromonas que ya son parcialmente solubles por sí mismas (aunque tienen un elevado peso molecular) [59, 68]. Por tanto, es probable que estas familias codifiquen proteínas genéricas de unión a ligando, con sólo algunos miembros involucrados en la quimiopercepción.

Proteínas quimiorreceptoras

Los quimiorreceptores están anclados en la membrana de las RN, e incluyen: la superfamilia de quimiorreceptores (que engloba los receptores olfativos y gustativos; ORs y GRs, respectivamente) [54, 55, 69], y la superfamilia de receptores ionotrópicos (comprendida por los IRs antenales y divergentes; aIRs y dIRs) [70].

Los análisis filogenéticos han demostrado que los ORs divergieron a partir de los GRs [69], de tal forma que todavía conservan ciertas similitudes estructurales y funcionales: (i) ambas familias multigénicas codifican para proteínas de unos 400 aminoácidos, caracterizadas por siete dominios transmembrana y una topología atípica (con el extremo N-terminal en el citosol y el C-terminal en la matriz extracelular [71, 72]); (ii) algunos GRs y ORs no son funcionales por sí mismos, sino que forman parte de complejos heteromultiméricos, como el dímero constituido por los Gr21a y Gr63a (de respuesta a CO_2) [73], o el que componen el co-receptor Or83b (también conocido como ORCO) y cualquier otro tipo de OR [74]. La superfamilia de receptores ionotrópicos también codifica para proteínas con función olfativa (los aIRs) y -probablemente- gustativa (los dIRs) [70, 75].

¿Por qué estas dos superfamilias multigénicas se han especializado independientemente (convergencia evolutiva) en funciones gustativas y olfativas? Las evidencias sugieren que en realidad discriminan compuestos de diferente naturaleza química. Así, los aIRs se expresan en sensilios coelónicos (asociados con la detección de aminas), mientras que los ORs en sensilios tricoideos (feromonas volátiles) y basicónicos (odorantes de nutrientes) [76]. Además, aunque no se conoce completamente su mecanismo de activación [51] (Figura 1.6), la cinética de ambos quimiorreceptores es diferente: el tiempo de reacción frente a estímulos es menor en OR-ORCO que en los aIRs (pero los aIRs prolongan el disparo durante más tiempo) [77, 78].

Con todo, se ha constatado que los IRs tienen un origen antiguo, y acometen funciones basales en la quimiopercepción [75]. Al contrario, el origen y evolución de los GRs y ORs parece vinculado a ciertas particularidades del linaje de los artrópodos [61]. Por ejemplo, la cinética de activación OR-ORCO (más sensible y rápida) se ha interpretado como una adaptación al vuelo, donde el tiempo de exposición al estímulo es limitante (la capacidad de volar fue adquirida por los insectos en el Carbonífero, hace unos 315-360 millones de años) [78, 79].

1.5 Evolución de las familias multigénicas del sistema quimiosensorial

1.5.1 Regiones codificadoras

Un aspecto fundamental para comprender el origen y función de estas familias multigénicas es cuantificar la contribución de la selección natural en su evolución. Esta información puede ser de gran utilidad para comprender si el impacto de la selección natural depende del proceso biológico en el que par-

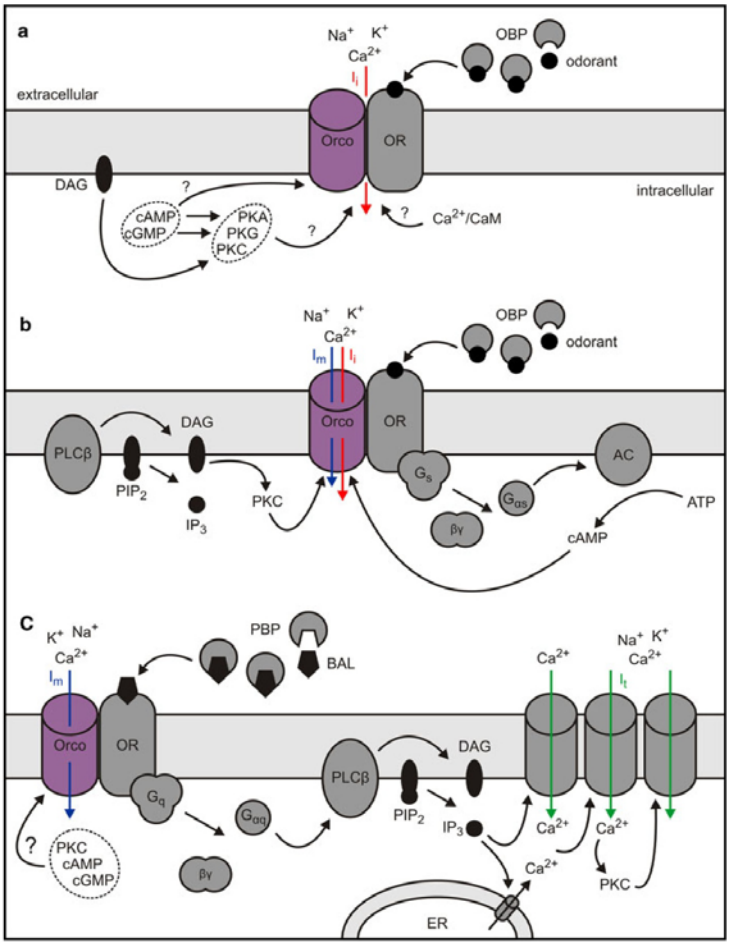


Figura 1.6: Las tres hipótesis sobre el mecanismo de activación del heterodímero OR-ORCO. A. Transducción meramente ionotrópica; B. Transducción tanto ionotrópica como metabotrópica; C. Transducción únicamente metabotrópica. Fuente: [51].

ticipan las proteínas (i.e. olfato y gusto) o de la función molecular que desempeñan (proteína de unión a ligando o quimiorreceptor).

Estas cuestiones se pueden abordar analizando los niveles de constricción funcional (ω) en sus regiones codificadoras. En el grupo *melanogaster* de *Drosophila* (Figura 1.7), la ω es mayor en las CSPs (la mediana es $\omega_{CSPs} = 0.05$), que en las familias del sistema olfativo ($\omega_{aIRs} = 0.12$, $\omega_{OBPs} = 0.14$ y $\omega_{ORs} = 0.14$) y gustativo ($\omega_{dIRs} = 0.18$ y $\omega_{GRs} = 0.22$) [80]. Como se ha comentado anteriormente, los valores de $\omega < 1$ indican que estas familias evolucionan -mayormente- bajo la acción de la selección purificadora. No obstante, se ha detectado la huella de la selección adaptativa en algunas posiciones de los ORs y, especialmente, de

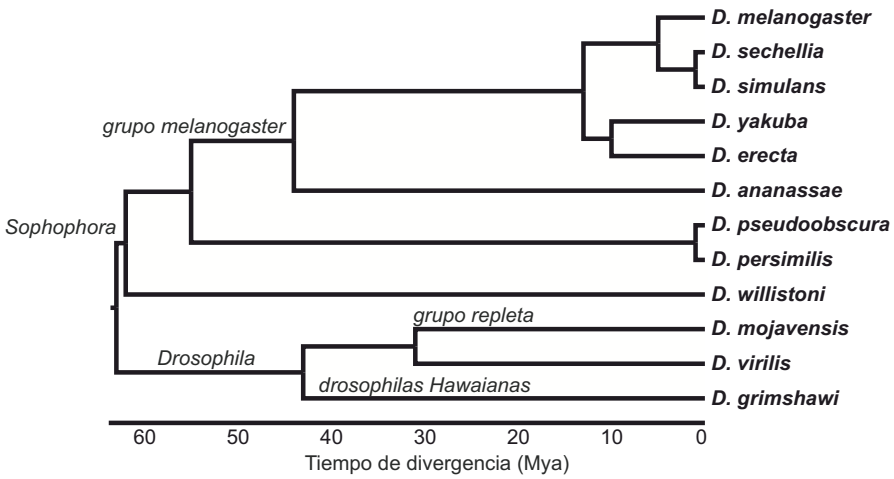


Figura 1.7: Relaciones filogenéticas de las 12 especies de *Drosophila*, con los tiempos de divergencia inferidos por Tamura [84].

los GRs que participan en la discriminación de sabores amargos [81, 82]. Muchos de los compuestos con sabor amargo, como la cafeína, son metabolitos secundarios producidos por las plantas, que actúan como insecticidas naturales ante los depredadores. Se hipotetiza que la presión selectiva que ejercen plantas e insectos entre sí (coevolución) es la causa de los recurrentes procesos adaptativos en este grupo de GRs [42, 83].

La ausencia de la huella de la selección positiva en las regiones codificadoras de proteínas de unión a ligando sugiere que los quimiorreceptores son los principales contribuyentes a los cambios adaptativos. No obstante, también es posible que las OBPs y CSPs tengan un papel adaptativo, pero a una escala temporal diferente. En este sentido, se ha inferido que la fijación de algunas mutaciones en el sitio de unión a ligando de las OBPs han sido promovidas por la selección natural positiva, tanto en la subfamilia Minus-C de *Apis mellifera* como en los duplicados recientes *Obp83a* y *Obp83b* de *D. melanogaster* [85, 86]. Además, las substituciones en las regiones codificadoras tan sólo representan una pequeña fracción de los posibles cambios fijados por la selección natural. Por tanto, para tener una visión general del impacto de la selección natural es fundamental estudiar otras regiones genómicas, y a diferentes tiempos evolutivos.

1.5.2 Regiones reguladoras de la transcripción

Los cambios en la regulación transcripcional pueden ser uno de los principales contribuyentes a la divergencia fenotípica. Ciertamente, la detección del estí-

mulo externo y la posterior respuesta conductual son procesos parcialmente independientes. Por un lado, la detección del estímulo depende de la activación del quimiorreceptor, mientras que la respuesta frente a ese estímulo del centro neuronal donde proyecta cada RN (ej. glomérulo olfativo). Es decir, el mismo quimiorreceptor expresado en dos RNs diferentes podría inducir respuestas conductuales dispares, incluso frente al mismo estímulo. Aunque todavía no se conoce completamente el mecanismo por el cual cada quimiorreceptor se transcribe -de forma específica- en un subconjunto de RNs, los análisis de expresión en ORs sugieren que la acción conjunta de unos pocos CREs proximales (que inducen transcripción basal) y distales (que reprimen la transcripción en las RNs donde el OR no debe expresarse) es suficiente para conferir un correcto patrón transcripcional [87, 88]. Bajo este sencillo mecanismo de regulación, las mutaciones que alteran la funcionalidad de los CREs producirán una gran variabilidad fenotípica, pudiendo ser uno de los principales sustratos de la selección natural.

Los estudios de asociación a escala genómica (GWAS) han confirmado que una parte significativa de la varianza en el comportamiento olfativo y gustativo se explica por mutaciones en regiones reguladoras de la transcripción [89, 90]. El caso más ilustrativo es la especialización de *D. sechellia* en *Morinda citrifolia*, una planta arbórea que produce metabolitos secundarios repelentes para la mayoría de especies del género *Drosophila*. *D. sechellia*, sin embargo, presenta una serie de modificaciones transcripcionales, incluida una inserción de 4bp en la región *upstream* del gen *Obp57e* [42] que alteran su comportamiento, haciendo que los metabolitos secundarios del fruto pasen de ser repelentes a atractivos.

1.5.3 Ganancia y pérdida de genes quimiosensoriales

El tamaño (número de genes) de las principales familias multigénicas del sistema quimiosensorial varía entre las 12 especies de *Drosophila* actualmente secuenciadas (Figura 1.7) [11]. La familia quimiosensorial con mayor varianza es la de los GRs, con un rango que va desde 50 copias en *D. virilis* y *D. mojavensis* hasta 68 en *D. willistoni*. Por el contrario, la familia más estable es la de los aIRs, con 17 copias en todas las especies excepto en *D. sechellia*, donde sólo se han identificado 16.

Cabe destacar que en genomas de baja cobertura, como los de *D. sechellia* (4.9X) y *D. persimilis* (4.1X), el número de genes anotados puede ser menor, pero no porque se hayan perdido a lo largo de la evolución, sino porque no se hayan secuenciado o ensamblado correctamente [91]. Por tanto, para evitar posibles sesgos en los análisis de ganancia y pérdida de genes, es recomendable excluir estas especies con genomas de mala calidad (Figura 1.8). En las 10 especies de *Drosophila* restantes, la tasa de nacimiento y muerte de genes (λ =

[eventos por millón de años y por copia]) indica que los aIRs y las CSPs han experimentado pocos cambios de tamaño (λ_{aIRs} = 0.0002 y λ_{CSPs} = 0.0004), mientras que las familias de los dIRs, ORs, OBPs y GRs son muy dinámicas (λ_{dIRs} = 0.0059, λ_{ORs} = 0.0062, λ_{OBPs} = 0.0069 y λ_{GRs} = 0.0113 respectivamente) [80]. Aunque el número de familias analizadas no es muy elevado, existe una correlación significativa entre ω y λ (correlación de Pearson; P = 0.0270). Esta asociación sugiere que la variación tanto a nivel de secuencia codificadora, como de número de copias, está moldeada por los mismos procesos evolutivos, ya sean selectivos o demográficos [81, 92].
La selección natural puede moldear tanto la ganancia como la pérdida de genes, dos tipos de eventos que pueden tener importantes consecuencias funcionales. Por ejemplo, la pseudogenización del quimiorreceptor *Gr64e* reduce la preferencia de *D. pseudoobscura* hacia el glicerol, un compuesto derivado de la fermentación por levadura (aunque la levadura es uno de los principales alimentos de *Drosophila*, *D. pseudoobscura* muestra una preferencia similar ante levadura y glucosa) [93]. En general, la especialización de nicho ecológico, ya sea en la dieta o en cualquier otro proceso vital, comporta una relajación de la selección natural (muchos genes son innecesarios y, por tanto, son susceptibles a perderse).
Los eventos demográficos también pueden tener un impacto en la variación del tamaño de las familias multigénicas. Así, en especies endémicas (donde el tamaño efectivo poblacional es pequeño), la eficacia de la selección natural es reducida (el efecto de la deriva genética es comparativamente grande), lo que puede incrementar la varianza en el número de copias [81]. En las 12 especies de *Drosophila*, se ha constatado que el endemismo (*D. sechellia* y *D. grimshawi*) y, en menor medida, el especialismo (*D. sechellia*, *D. erecta* y *D. mojavensis*) tienen efectos significativos en el número de quimiorreceptores (ORs y GRs), pero no en el de proteínas de unión a ligando (OBPs y CSPs) (Figura 1.8) [92].

1.5.4 Distribución genómica de las OBPs

Las principales mutaciones estructurales de los cromosomas son las inversiones paracéntricas, causadas -por ejemplo- por eventos de recombinación ectópica entre secuencias repetitivas e invertidas (cercanas a los puntos de rotura) [94]. En *Drosophila*, asumiendo los tiempos de divergencia inferidos por Tamura [84], las tasas de reordenamiento cromosómico van desde 0.015 hasta 0.093 puntos de rotura por millón de años (Mya) y por megabase [95], lo que aproximadamente equivale a 0.187-1.157 inversiones fijadas por Mya y por cromosoma. Teniendo en cuenta que 60 inversiones (producidas de forma aleatoria) son suficientes para cambiar la localización de la mayoría de genes de un cromosoma [96], la colinearidad genómica debería mantenerse poco, incluso entre especies cercanas.

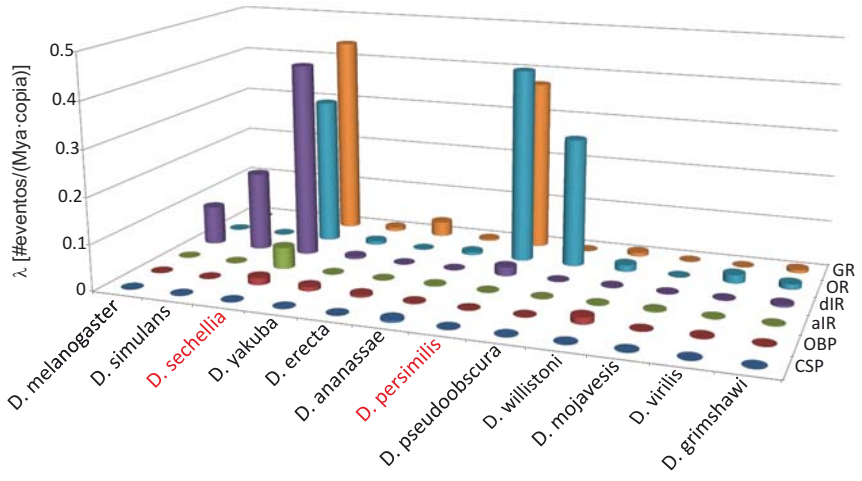


Figura 1.8: Tasa de nacimiento y muerte de genes (λ) por familia y especie. Nótese como las especies con un genoma de mala calidad (en rojo) tienen λ mayores.

Entonces, ¿cómo se explica que algunos grupos de genes conserven su 'vecindad' (se mantengan contiguos) a lo largo del tiempo, formando clusters? En *Drosophila*, la mayor parte de la distribución génica se explica por un sesgo mutacional: como los puntos de rotura no están distribuidos aleatoriamente a lo largo del cromosoma, algunas regiones son menos propensas a experimentar reordenaciones génicas. Aun así, existen clusters activamente mantenidos por la selección natural [97]. Para comprender el significado biológico de estos clusters, primero es imprescindible discernirlos de forma inequívoca, contrastando si realmente están más conservados de lo esperado dados los propios niveles y patrones de reordenamiento cromosómico.

Los genes que codifican OBPs -en particular- no están distribuidos aleatoriamente en el genoma (Figura 1.9). Dicha distribución podría explicarse por su principal mecanismo de origen (la duplicación por entrecruzamiento desigual), que lleva a una disposición en tándem de las copias génicas. Sin embargo, se ha comprobado que los genes que codifican para OBPs mantienen su posición física más tiempo del esperado, dadas las tasas de reordenación cromosómica de *Drosophila* [61]. Esto sugiere que la distribución tiene algún significado funcional, como por ejemplo facilitar la co-regulación con otros genes, ya sean parálogos o no.

La localización de los genes puede estar restringida por los mecanismos de co-regulación transcripcional (ej. CREs compartidos entre genes o estado de la cromatina). Ciertamente, las mutaciones que alteran la localización de los genes pueden inducir un patrón de co-expresión erróneo y, por tanto, tende-

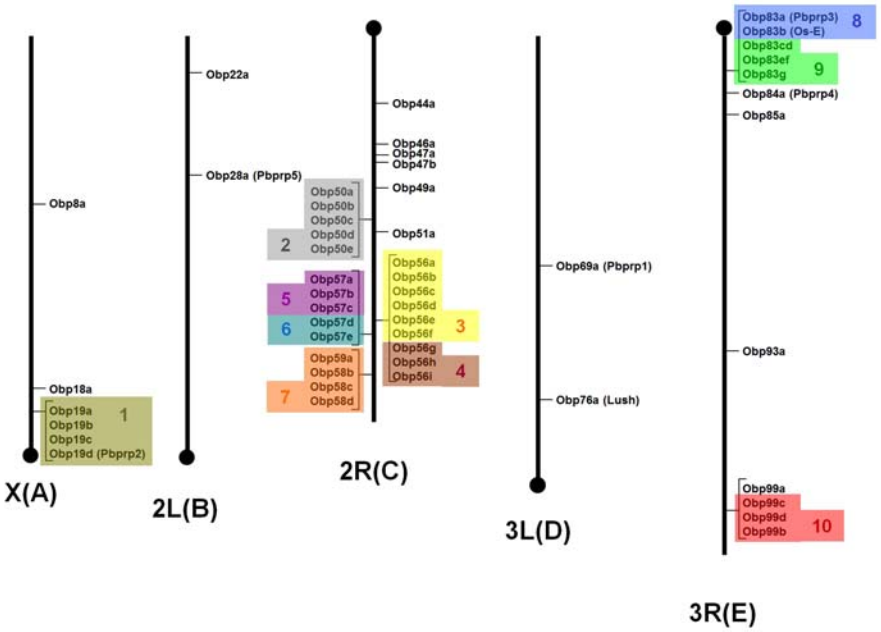


Figura 1.9: Distribución de los genes que codifican OBPs a lo largo de los principales brazos cromosómicos de *D. melanogaster*. El sombreado de colores son los 10 clusters inferidos por [98].

rán a ser eliminadas por la selección natural (lo que dejará una huella molecular característica en la distribución génica). Por ejemplo, en insectos, los genes co-regulados se mantienen físicamente cercanos por estar dentro del rango de acción de sus *enhancers* [99].

Como los dominios de la cromatina proporcionan diferentes ambientes transcripcionales, también pueden afectar a la organización génica. Por ejemplo, Caron y colaboradores observaron que los genes con altos niveles de expresión (EI o *expression intensity*) están físicamente agrupados en el genoma humano [100]. En realidad, dicha relación es artefactual, ya que los genes agrupados son los que tienen una elevada amplitud transcripcional (EB o *expression breadth*: número de tejidos o condiciones donde el gen se transcribe). No obstante, como EB y EI están positivamente asociados, también se observa el efecto indirecto de EI [101].

¿Por qué los genes con alto EB suelen mantener su proximidad cromosómica? Se ha hipotetizado que la distribución no aleatoria de los genes es resultado de presiones selectivas para minimizar el ruido transcripcional (EN o *expression noise*, cambios en la abundancia de transcrito que no son debidos a diferencias genéticas ni ambientales, sino a la propia estocasticidad del proceso transcripcional). El EN es -normalmente- deletéreo, porque genera importantes des-

ajustes estequiométricos en las redes metabólicas [102, 103]. Este desajuste es mayor para los genes con alto EB (coloquialmente denominados genes '*housekeeping*'), ya que normalmente ocupan posiciones centrales en las vías metabólicas. Por tanto, la selección natural tiende a favorecer la localización de los genes con alto EB en regiones cromosómicas de bajo EN.
En este contexto, las proteínas que regulan la organización de la cromatina pueden tener un papel relevante, especialmente las que están -directa o indirectamente- asociadas a la membrana nuclear [104-108]. La membrana nuclear, además de compartimentalizar la célula, determina la posición y la conformación de los cromosomas en el nucleoplasma. Entre otras funciones, dirige la acción de las proteínas remodeladoras de la cromatina, posibilitando el acceso de la maquinaria celular (proteínas involucradas tanto en la transcripción como en el transporte de macromoléculas). En *Drosophila*, una de las evidencias más contundentes acerca del papel de la membrana nuclear en la distribución génica es que los sitios de unión al ADN de lam_{D0} (proteína de la membrana nuclear) co-localizan con clusters ultraconservados [109]. A pesar de todos estos estudios, todavía se desconocen las causas y mecanismos que explican la distribución cromosómica de los genes que codifican OBPs.

1.6 Especies y poblaciones analizadas

1.6.1 Las 12 especies de *Drosophila*

Los insectos del género *Drosophila* son típicamente pequeños (2-8 milímetros de longitud), y se alimentan de fruta madura o en descomposición. Aunque la sistemática y nomenclatura del género todavía no está completamente resuelta, se han descrito tres grandes subgéneros: las *Idiomyia o Hawaianas* (unas 1000 especies), las *Sophophora* (unas 1100) y las *Drosophila* (unas 330) (Figura 1.7).
Desde hace más de un siglo, las especies del género *Drosophila* y, más concretamente *D. melanogaster*, son organismos modelo en varias disciplinas científicas, como en ecología, genética del desarrollo, evolutiva y molecular. Los motivos de su éxito como organismo modelo son: (i) la facilidad para mantenerlas en el laboratorio (son pequeñas y tienen una dieta sencilla), (ii) un tiempo de generación corto (10-20 días, dependiendo de la temperatura), (iii) una descendencia abundante (fundamental para testar hipótesis genéticas), (iv) y que presentan cromosomas politénicos (esenciales para estudios cromosómicos).
La secuenciación de *D. melanogaster* [110] y, posteriormente, de *D. pseudoobscura* [111] establecieron las bases para realizar estudios evolutivos impensables hasta la fecha. Con la secuenciación de otros 10 genomas de *Drosophila* (*D. sechellia*, *D. simulans*, *D. yakuba*, *D. erecta*, *D. ananassae*, *D. persimilis*, *D. mojavensis*, *D. willistoni*, *D. virilis* y *D. grimshawi*) se completó uno de los grandes

hitos de la genómica comparada en organismos eucariotas (Figura 1.7) [11]. Hoy en día, con la creciente disponibilidad de datos masivos funcionales (ej. expresión génica [65, 112], recombinación [113], anotación de CREs [112]), las 12 especies de *Drosophila* son uno de los mejores modelos de estudio en genómica evolutiva.

1.6.2 *Drosophila Genetic Reference Panel*

El *Drosophila Genetic Reference Panel* (DGRP) es un conjunto de 168 líneas isogénicas de *D. melanogaster*, y de sus respectivas secuencias genómicas [114]. Para construir las líneas isogénicas, se partió de hembras fecundadas, y se realizaron cruzamientos hermano-hermana durante un total de 20 generaciones. Posteriormente, 10 de las líneas se secuenciaron con la plataforma 454 de Roche (con una cobertura media de 12.1X), 129 con Illumina (21.4X) y 29 con ambas. El DGRP se concibió como un recurso para realizar estudios de asociación genotipo-fenotipo (GWAS). No obstante, la secuencia genómica de 168 líneas de una misma población (Raleigh, Carolina del Norte, Estados Unidos de América) también posibilita realizar inferencias evolutivas a una resolución sin precedentes, incluyendo la posibilidad de detectar procesos incipientes de adaptación a nivel molecular.

2
Objetivos

Todos los organismos presentan sistemas capaces de discriminar compuestos con gran especificidad y sensibilidad. Incluso las bacterias pueden detectar la concentración de ciertas substancias, como la glucosa, y condicionar su localización espacial a dicho gradiente, a través de un fenómeno conocido como *quimiotaxis* [115]. La ubicuidad del sistema quimiosensorial (SQ) refleja su esencialidad e importancia en la eficacia biológica de los individuos, lo que le convierte en un excelente modelo para estudiar procesos adaptativos a nivel molecular.

El principal objetivo de esta tesis doctoral es analizar el impacto de la selección natural, tanto positiva (o *darwiniana*) como negativa (o purificadora), en la regulación transcripcional de los genes del SQ. Este objetivo general se ha abordado mediante: (i) el desarrollo e implementación de herramientas bioinformáticas que faciliten el análisis de la variabilidad genética; (ii) el estudio del impacto de la selección natural en los mecanismos de regulación transcripcional de los genes del SQ de *Drosophila* (regiones *upstream* y dominios de la cromatina). Los objetivos específicos de la tesis doctoral han sido:

- Implementar nuevos métodos de genética de poblaciones y evolución molecular en el programa DnaSP [116], para: (i) analizar la variabilidad genética en *loci* independientes, (ii) detectar regiones conservadas a lo largo del tiempo (*phylogenetic footprinting* y *phylogenetic shadowing*), y (iii) visualizar los resultados de DnaSP conjuntamente con las anotaciones existentes en el navegador genómico de la Universidad de California y Santa Cruz (UCSC) [117].

- Implementar una instancia del navegador genómico *GBrowse* [118] que permita calcular y visualizar los niveles y patrones de variabilidad nucleotídica en las líneas de *D. melanogaster* secuenciadas por el proyecto DGRP.

- Desarrollar e implementar modelos estocásticos ganancia y pérdida de elementos genéticos (ej. genes o CREs) que posibiliten testar hipótesis biológicamente relevantes, como expansiones y contracciones en especies concretas.
- Estudiar el impacto de diferentes mecanismos de co-regulación transcripcional en la distribución cromosómica de los genes que codifican *Odorant-Binding Proteins* (OBPs).
- Examinar la contribución de la selección natural a la evolución de las regiones *upstream* de los genes del SQ de *D. melanogaster*, con el propósito de determinar su potencial papel en la adaptación de esta especie durante la colonización de nuevos hábitats.

3
Capítulos

3.1 DnaSP v5: a software for comprehensive analysis of DNA polymorphism data

DnaSP es un programa que permite analizar de forma exhaustiva el polimorfismo nucleotídico. La versión 5 implementa nuevas características y métodos analíticos especialmente diseñados para estudiar el polimorfismo de ADN en datos masivos. Entre otras características, las nuevas implementaciones posibilitan: (i) analizar varios archivos de datos a la vez; (ii) reconstruir la fase haplotípica de individuos diploides; (iii) examinar el nivel y patrón de inserción y deleción de nucleótidos; (iv) visualizar los resultados conjuntamente con las anotaciones genómicas disponibles en la navegador genómico de la Universidad de California y Santa Cruz (UCSC); (v) inferir regiones funcionales mediante metodologías de *phylogenetic footprinting*.
A diferencia de las regiones codificadoras de proteína, la anotación de elementos *cis*-reguladores (CREs) de la transcripción sigue siendo un importante reto metodológico. La capacidad de inferir CREs difiere sustancialmente *in vitro* e *in vivo*. *In vivo*, además de la secuencia de nucleótidos, hay otros determinantes de la unión proteína-ADN, como el estado de la cromatina o la topología del ADN [119, 120]. En este contexto, las metodologías de *phylogenetic footprinting* y *phylogenetic shadowing* han demostrado ser de gran utilidad para la detección de regiones funcionales, principalmente CREs.

BIOINFORMATICS APPLICATIONS NOTE Vol. 25 no. 11 2009, pages 1451–1452
doi:10.1093/bioinformatics/btp187

Genetics and population analysis

DnaSP v5: a software for comprehensive analysis of DNA polymorphism data

P. Librado[1,2] and J. Rozas[1,2,*]

[1]Departament de Genètica, Facultat de Biologia and [2]Institut de Recerca de la Biodiversitat, Universitat de Barcelona, Diagonal 645, 08028 Barcelona, Spain

Received on February 10, 2009; revised and accepted on April 2, 2009

Advance Access publication April 3, 2009

Associate Editor: Martin Bishop

ABSTRACT

Motivation: DnaSP is a software package for a comprehensive analysis of DNA polymorphism data. Version 5 implements a number of new features and analytical methods allowing extensive DNA polymorphism analyses on large datasets. Among other features, the newly implemented methods allow for: (i) analyses on multiple data files; (ii) haplotype phasing; (iii) analyses on insertion/deletion polymorphism data; (iv) visualizing sliding window results integrated with available genome annotations in the UCSC browser.

Availability: Freely available to academic users from: http://www.ub.edu/dnasp

Contact: jrozas@ub.edu

1 INTRODUCTION

The analysis of DNA polymorphisms is a powerful approach to understand the evolutionary process and to establish the functional significance of particular genomic regions (Begun *et al.*, 2007; Nielsen, 2005; Rosenberg and Nordborg, 2002). In this context, estimating the impact of natural selection (both positive and negative) is of major interest. Furthermore, DNA polymorphisms are relevant as a tool for a broad range of life science disciplines. Consequently, many high-throughput sequencing, genotyping and polymorphism detection systems have been developed and are currently publicly available (Shendure and Ji, 2008). These new technologies are generating massive amounts of data that need to be processed, analyzed and transformed effectively into knowledge.

These technological advances have largely stimulated the development of both analytical methods and computer applications. Population genetic methods, and particularly those based on coalescent theory (Hudson, 1990; Wakeley, 2009), are used at an increasing rate, but need to be adapted to the particularities of the data (massive amounts of data, missing data, genotypes, insertion/deletion (indels) polymorphisms, etc.). Furthermore, new computer applications and algorithms need to be developed for processing massive datasets (Excoffier and Heckel, 2006), and more specifically computer visualization tools for the representation of DNA variation patterns. DnaSP (DNA Sequence Polymorphism) is a software package that allows for extensive DNA polymorphism analyses using a friendly graphical user interface (GUI) (Rozas *et al.*, 2003). Version 5 extends the capabilities of the software, allowing comprehensive DNA polymorphism analyses on multiple data files and on large datasets. Altogether, the present version of DnaSP has the appropriate features for exhaustive exploratory analyses using high-throughput DNA polymorphism data.

*To whom correspondence should be addressed.

2 FEATURES

DnaSP v5 incorporates major improvements. The new version currently allows for the handling and analysis of multiple data files in batch, and implements new algorithms and methods; among other things (see below) includes a new module to identify conserved DNA regions, this feature might be useful for phylogenetic footprinting-based analyses (Vingron *et al.,* 2009). DnaSP provides a convenient GUI facilitating all data management and analytical tasks; the results can be visualized graphically as well as in a text report. DnaSP accepts multiple DNA sequence alignment file formats (Rozas *et al.*, 2003), including NEXUS (Maddison *et al.*, 1997), and HapMap3 files with phased haplotypes (The International HapMap Consortium, 2003). The software allows exhaustive DNA polymorphism analyses, including those based on coalescent theory (Rozas *et al.*, 2003; Wakeley, 2009).

2.1 Haplotype reconstruction

Haplotype reconstruction aims at resolving haplotype phase given genotypic information. DnaSP implements statistical methods to infer haplotype phase, and prepares adequately the phased data for subsequent analyses. The input data (unphased genotype data) are required in FASTA format using IUPAC nucleotide ambiguity codes to represent heterozygous sites. DnaSP reconstructs the phase by applying various algorithms (PHASE v2.1, fastPHASE v1.1 and HAPAR) differing in the underlying population genetic assumptions. PHASE (Stephens and Donnelly, 2003; Stephens *et al.*, 2001) assumes Hardy–Weinberg equilibrium and uses a coalescent-based Bayesian method to infer haplotypes. fastPHASE (Scheet and Stephens, 2006) implements a modification of the PHASE algorithm taking into account the patterns of linkage disequilibrium and its gradual decline with physical distance. This algorithm is faster and allows for the handling of larger datasets than PHASE, while being slightly less accurate. HAPAR (Wang and Xu, 2003) infers haplotype phase by maximum parsimony, i.e. attempts to find the minimum number of haplotypes explaining the genotype sample.

2.2 Deletion/insertion polymorphisms

Deletion/insertion polymorphisms (DIPs) analysis can provide insights into the evolutionary forces acting on DNA. This information, however, has been rarely used. One obstacle has been the difficulty of defining clearly homologous states (Young and Healy, 2003). DnaSP incorporates an algorithm for treating indels related to the 'simple indel coding' method of Simmons and Ochoterena (2000). Specifically, only indels with the same 5′ and 3′ termini are considered homologous (resulted from a single event), and indels of different lengths (even in the same position of the alignment) are treated as different events. DnaSP, nevertheless, uses a slightly different method for coding completely overlapping gaps, and allows the user to choose the level of overlap to be coded. Subsequently, DnaSP estimates a number of DIP summary statistics, such as the average indel length, indel diversity, as well as Tajima's *D* (Tajima, 1989) based on indel information. Additionally, it exports the recoded data in the NEXUS format file.

2.3 Analysis of multiple data files

DnaSP can automatically read and analyze multiple data files sequentially (in batch mode). These data files may contain a varying number of sequences (from within one species, or from one species as well as one outgroup), or represent diverse genomic regions. The program estimates the most common DNA polymorphism and divergence summary statistics (such as the nucleotide and haplotype diversity, the population mutation parameter, the number of nucleotide substitutions per site, etc.), and neutrality tests (such as Tajima's, Fu and Li's and Fu's tests).

2.4 Sliding window results visualization

The sliding window technique is a useful tool for exploratory DNA polymorphism data analysis (Hutter *et al.*, 2006; Rozas *et al.*, 2003; Vilella *et al.*, 2005). The current version of DnaSP permits visualizing results of the sliding window (for example, nucleotide diversity or Tajima's *D* values along the DNA sequence) integrating available genome annotations in the UCSC browser (Kent *et al.*, 2002). This feature can greatly facilitate the interpretation of the results; for instance, it is possible to identify the relevant genome annotations (genes, intergenic regions, conserved regions, etc.), which are adjacent to regions with atypical patterns of nucleotide variation.

3 IMPLEMENTATION

DnaSP version 5 has been developed in Microsoft Visual Basic v6.0, C and C++, and it runs under Microsoft Windows operating systems (2000/XP/Vista). With the use of Windows emulators, DnaSP can also run on Apple Macintosh platforms, Linux and Unix-based operating systems. The software has been tested in all three platforms.

ACKNOWLEDGEMENTS

We acknowledge Sergios-Orestis Kolokotronis for helpful comments on the manuscript. Special thanks to the numerous users who tested the software with their data, and particularly to all members of the Molecular Evolutionary Genetics group at the Departament de Genètica, Universitat de Barcelona.

Funding: Spanish Dirección General de Investigación Científica y Técnica (grants BFU2004-02253 and BFU2007-62927); the Catalonian Comissió Interdepartamental de Recerca i Innovació Tecnològica (grant 2005SGR00166).

Conflict of Interest: none declared.

REFERENCES

Begun,D.J. *et al.* (2007) Population genomics: whole-genome analysis of polymorphism and divergence in Drosophila simulans. *PLoS Biol.*, **6**, e310.

Excoffier,L. and Heckel,G. (2006) Computer programs for population genetics data analysis: a survival guide. *Nat. Rev. Genet.*, **7**, 745–758.

Hudson,R.R. (1990) Gene genealogies and the coalescent process. *Oxf. Surv. Evol. Biol.*, **7**, 1–44.

Hutter,S. *et al.* (2006) Genome-wide DNA polymorphism analyses using VariScan. *BMC Bioinformatics*, **7**, 409.

Kent,W.J. *et al.* (2002) The Human Genome Browser at UCSC. *Genome Res.*, **12**, 996–1006.

Maddison,W.P. *et al.* (1997) NEXUS: an extendible file format for systematic information. *Syst. Biol.*, **46**, 590–621.

Nielsen,R. (2005) Molecular signatures of natural selection. *Annu. Rev. Genet.*, **39**, 197–218.

Rosenberg,N.A. and Nordborg,M. (2002) Genealogical trees, coalescent theory, and the analysis of genetic polymorphisms. *Nat. Rev. Genet.*, **3**, 380–390.

Rozas,J. *et al.* (2003) DnaSP, DNA polymorphism analyses by the coalescent and other methods. *Bioinformatics*, **19**, 2496–2497.

Scheet,P. and Stephens,M. (2006) A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *Am. J. Hum. Genet.*, **78**, 629–644.

Shendure,J. and Ji,H. (2008) Next-generation DNA sequencing. *Nat. Biotechnol.*, **26**, 1135–1145.

Simmons,M.P. and Ochoterena, H. (2000) Gaps as characters in sequence-based phylogenetic analyses. *Syst. Biol.*, **49**, 369–381.

Stephens,M. and Donnelly,P. (2003) A comparison of Bayesian methods for haplotype reconstruction from population genotype data. *Am. J. Hum. Genet.*, **73**, 1162–1169.

Stephens,M. *et al.* (2001) A new statistical method for haplotype reconstruction from population data. *Am. J. Hum. Genet.*, **68**, 978–989.

Tajima,F. (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, **123**, 585–595.

The International HapMap Consortium (2003) The International HapMap Project. *Nature*, **426**, 789–796.

Vilella,A.J. *et al.* (2005) VariScan: analysis of evolutionary patterns from large-scale DNA sequence polymorphism data. *Bioinformatics*, **21**, 2791–2793.

Vingron,M. *et al.* (2009) Integrating sequence,evolution and functional genomics in regulatory genomics. *Genome Biol.*, **10**, 202.

Wang,L. and Xu,Y. (2003) Haplotype inference by maximum parsimony. *Bioinformatics*, **19**, 1773–1780.

Wakeley,J. (2009) *Coalescent Theory. An Introduction*. Roberts and Company Publishers. Greenwood Village.

Young,N.D. and Healy, J. (2003) GapCoder automates the use of indel characters in phylogenetic analysis. *BMC Bioinformatics*, **4**, 6.

3.2 BadiRate: estimating family turnover rates by likelihood-based methods

El análisis comparativo de especies filogenéticamente relacionadas permite inferir el papel de la selección natural y la adaptación en la evolución de las familias multigénicas. Sin embargo, los métodos existentes no eran apropiados para estudiar la dinámica de ciertos tipos de elementos genéticos funcionales, tales como los elementos *cis*-reguladores de transcripción (CREs).
En este artículo se describe BadiRate, un nuevo programa para estimar tanto la dinámica de familias de elementos funcionales (ej. genes o CREs), como su contenido en los nodos internos de la filogenia. BadiRate implementa varios modelos estocásticos, los cuales proporcionan un marco estadístico apropiado para contrastar hipótesis, como expansiones y contracciones de familias multigénicas en linajes específicos. Las pruebas de validación (en datos simulados y empíricos) muestran que BadiRate es capaz de inferir las tasas de ganancia y pérdida de elementos funcionales con gran exactitud.

BIOINFORMATICS APPLICATIONS NOTE

Vol. 28 no. 2 2012, pages 279–281
doi:10.1093/bioinformatics/btr623

Genome analysis

Advance Access publication November 10, 2011

BadiRate: estimating family turnover rates by likelihood-based methods

P. Librado, F. G. Vieira and J. Rozas*

Departament de Genètica and Institut de Recerca de la Biodiversitat (IRBio), Universitat de Barcelona, Diagonal 645, 08028 Barcelona, Spain

Associate Editor: David Posada

ABSTRACT

Motivation: The comparative analysis of gene gain and loss rates is critical for understanding the role of natural selection and adaptation in shaping gene family sizes. Studying complete genome data from closely related species allows accurate estimation of gene family turnover rates. Current methods and software tools, however, are not well designed for dealing with certain kinds of functional elements, such as microRNAs or transcription factor binding sites.

Results: Here, we describe BadiRate, a new software tool to estimate family turnover rates, as well as the number of elements in internal phylogenetic nodes, by likelihood-based methods and parsimony. It implements two stochastic population models, which provide the appropriate statistical framework for testing hypothesis, such as lineage-specific gene family expansions or contractions. We have assessed the accuracy of BadiRate by computer simulations, and have also illustrated its functionality by analyzing a representative empirical dataset.

Availability: BadiRate software and documentation is available from http://www.ub.edu/softevol/badirate.

Contact: jrozas@ub.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on July 24, 2011; revised on November 4, 2011; accepted on November 7, 2011

1 INTRODUCTION

It is generally accepted that gene and genome duplications are a major evolutionary mechanism for generating functional innovation (Ohno, 1970). The increasing availability of closely related genome sequences allows an accurate analysis of gene family evolution (Hahn *et al*., 2007; Sanchez-Gracia *et al*., 2009; Vieira and Rozas, 2010). Such studies have shown that most families are highly dynamic and evolve under a birth-and-death (BD) process (Nei and Rooney, 2005). Indeed, the comprehensive analysis of gene gains and losses can provide helpful insight into the role of natural selection and adaptation in shaping gene family size variation.

The stochastic BD model (BDM) (Hahn *et al*., 2005) implemented in the programs CAFE (De Bie *et al*., 2006) and BEGFE (Liu *et al*., 2011) allows estimating the family turnover rate (λ) by maximum likelihood (ML) and by Bayesian methods, respectively. This model, nevertheless, has some drawbacks. First, it assumes equal BD rates, an assumption that may not hold. Secondly, because duplications from zero ancestral genes are not possible (zero is an absorbing state in the probabilistic BDM), it cannot handle gene families without elements in the phylogenetic root. These assumptions can therefore bias the estimates of both the number of members in internal nodes, as well as the BD rate. Two recently developed computer programs, GLOOME (Cohen *et al*., 2010) and Count (Csuros, 2010), overcome these difficulties. Nevertheless, GLOOME can only model presence/absence of phyletic patterns instead of size changes, whereas Count assumes independent turnover rates for all lineages, which precludes testing biological relevant hypothesis such as linage-specific accelerations.

Here, we describe BadiRate, a new software tool to estimate family turnover rates through the Gain-and-Death (GD) and the Birth-Death-and-Innovation (BDI, also known as Birth-Death-and-Immigration) stochastic models. The current implementation allows modeling families of diverse functional elements, such as microRNAs, *cis*-regulatory elements or coding-protein genes. Additionally, these models provide a statistical framework for hypothesis testing, such as family expansions/contractions in specific lineages.

2 DESCRIPTION

BadiRate implements methods to estimate the family turnover rates such as, gain (γ), birth (β), death (δ) and innovation (ι) rates. These rates can be classified as density-dependent (BD) and density-independent (gain and innovation). Indeed, the probability of having a death (e.g. a gene loss via deletion or pseudogenization) or a birth (e.g. a gene gain by unequal crossing-over) event is proportional to the actual family size. Conversely, the probability of having a gain or innovation [e.g. horizontal gene transfer (HGT) or by *de novo* origin of short *cis*-regulatory elements] event does not depend on the actual gene number.

We have implemented two stochastic population models in BadiRate, the BDI and the GD models (Csuros and Miklos, 2009; Hahn *et al*., 2005) (for details see Supplementary Material), to analyze the evolution of such diverse functional elements. The GD model is especially suitable for analyzing families where members might have a *de novo* origin, such as transcription factor binding sites (TFBSs) and small non-coding RNAs (miRNAs, piRNAs, etc) or even families exhibiting high HGT. In contrast, the BDI model is appropriated to study gene families whose major mechanism for the acquisition of genes is density-dependent (although it can also model scenarios with a reduced number of HGT or *de novo* origin events). BadiRate also implements a particular case of the BDI model in which BD rates are assumed to be equal, the Lambda-Innovation

*To whom correspondence should be addressed.

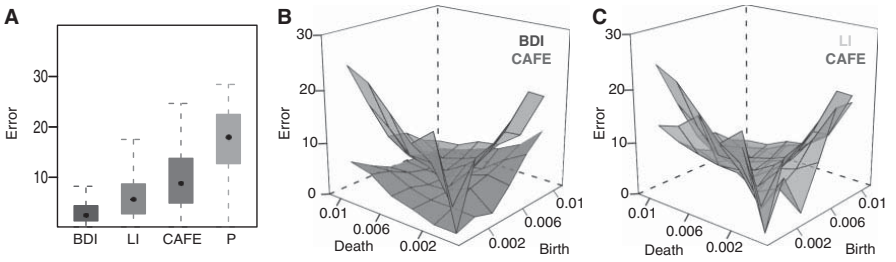


Fig. 1. Relative performance of the different methods. Computer simulations based on gene families of five members ($S = 5$) at the most internal node, and a null innovation rate ($I = 0$). (**A**) The box plot represents the error values (see Supplementary Material) distribution across tested methods. (**B** and **C**) Surface plots showing the error values of the ML estimates in scenarios simulated under a combination biologically realistic BD rates. Analysis under the BDI, LI and P models are conducted with BadiRate and are depicted in blue, green and cyan, respectively. The model implemented in CAFE is depicted in red. BD rates are measured in number of events per gene per million years (See Supplementary Material for details).

model (LI). This model is nearly equivalent to that implemented in CAFE (except for the innovation parameter), which allows the comparison between the two programs.

We have implemented three statistical frameworks to estimate family turnover rates and the number of members in internal nodes: maximum a posteriori (MAP), ML and parsimony (see Supplementary Material). Likelihood-based methods have the advantage of contrasting biological relevant scenarios, such as the identification of gene families or specific-lineages with extreme turnover rates (Johnson and Omland, 2004). Moreover, the MAP approach allows the incorporation of prior biological information without a large computational cost.

BadiRate requires as input the established species phylogenetic tree and a tab-delimited file with the family size of each species represented in the phylogeny (see BadiRate's documentation).

3 SIMULATION RESULTS

We assessed the accuracy of the turnover rates estimates by computer simulations on the well-characterized 12 *Drosophila species* phylogeny (Supplementary Material; Supplementary Figures S1, S2, S3, S4 and S5). Particularly, we benchmarked the BadiRate ML (under the BDI, LI and GD models) and parsimony estimates, as well as the CAFE (v2.2) ML estimates (under the implemented BDM model).

Our results show that, in general (and as expected), the parsimony algorithm performs worse than ML methods (Fig. 1A). Among the ML models, the BDI method outperforms all others (LI and CAFE), especially in cases where small-size families have asymmetric birth/death rates, i.e. $\beta > \delta$ or vice versa (Fig. 1B; Supplementary Figure S2). Apart from the higher accuracy of the BDI model, which mainly results from the separate estimation of BD rates, the LI model also outperforms CAFE even in scenarios with a null innovation rate (Fig. 1A and C; Supplementary Figure S3). Unlike LI and CAFE, which are particular cases of the BDI model, the performance of the GD model is not directly comparable to BDI. Still, our simulations show good performance of the GD model in the analyzed scenarios (Supplementary Figure S4).

4 EMPIRICAL RESULTS

We also illustrate the BadiRate application by analyzing the suggested miRNA expansion in the *D.willistoni* lineage (Nozawa *et al*., 2010). Since the identification of the miRNA copies in the 11 *Drosophila* species was conducted by similarity based on the available *D.melanogaster* miRNA data, the identification of the family members is less accurate for longer divergence times. To control for this effect, we contrasted two scenarios, one assuming independent turnover rates in two classes of branches (in the internal lineages leading to *D.melanogaster* and in the rest of branches) and the other incorporating a third class of turnover rates (for the *D.willistoni* lineage). The lower Akaike Information Criterion (AIC) value of the second scenario (AIC = 1509.8128) compared with the first one (AIC = 1518.3918) suggests that the *D.willistoni* lineage indeed has distinct miRNA turnover rates. We also inferred the most likely number of miRNA elements in the internal nodes of the *Drosophila* phylogeny; these figures are very similar to that estimated in (Nozawa *et al*., 2010) (Supplementary Figure S6).

ACKNOWLEDGEMENTS

We thank F.C. Almeida, M.C. Frias-Lopez, S. Guirao-Rico, A. Sanchez-Gracia and V. Soria-Carrasco, and three anonymous reviewers for their comments and suggestions on the manuscript and on the software.

Funding: This work was supported by grants from the Ministerio de Ciencia e Innovación of Spain (BFU2007-62927 and BFU2010-15484) and from the Comissió Interdepartamental de Recerca i Innovació Tecnològica of Spain (2009SGR-1287). J.R. was partially supported by ICREA Academia (Generalitat de Catalunya).

Conflict of Interest: none declared.

REFERENCES

Cohen,O. *et al.* (2010) GLOOME: gain loss mapping engine. *Bioinformatics*, **26**, 2914–2915.

Csuros,M. (2010) Count: evolutionary analysis of phylogenetic profiles with parsimony and likelihood. *Bioinformatics*, **26**, 1910–1912.

Csuros,M. and Miklos,I. (2009) Streamlining and large ancestral genomes in Archaea inferred with a phylogenetic birth-and-death model. *Mol. Biol. Evol.*, **26**, 2087–2095.

Supplementary Material

BadiRate: Estimating Family Turnover Rates by Likelihood-Based Methods

P. Librado, F. G. Vieira and J. Rozas
Departament de Genètica and Institut de Recerca de la Biodiversitat,
Universitat de Barcelona, Diagonal 643, 08028 Barcelona, Spain.

Supplementary Methods

Maximum Likelihood (ML) estimates

BadiRate implements an Amoeba hill-climbing method (Nelder and Mead, 1965) to maximize the likelihood of the family turnover rates:

$$\theta_{ML} = \arg\max_{\theta} L(\theta) \quad (1)$$

where θ are the family turnover rate parameters, and $L(\theta)$ is the likelihood of these parameters in the whole phylogeny. BadiRate assumes that the evolution of the families is independent, which allows computing the $L(\theta)$ as the product of the individual family likelihoods:

$$L(\theta) = \prod_{i=1}^{f} L_i(\theta) \quad (2a)$$

where f is the total number of families, and $L_i(\theta)$ is the likelihood of the parameters in the observed family i. Nevertheless, if some families became extinct in all surveyed species the actual value of f can be larger than the observed one. In such cases, it is convenient to correct the $L(\theta)$ value as (Cohen, et al., 2008; Felsenstein, 1992):

$$L(\theta) = \frac{\prod_{i=1}^{f} L_i(\theta)}{\left[\, 1 - L_{NULL}(\theta) \,\right]^{f}} \quad (2b)$$

where $L_{NULL}(\theta)$ is the likelihood of a family to be absent in all external nodes. BadiRate calculates $L_i(\theta)$ and $L_{NULL}(\theta)$ using the pruning algorithm of Felsenstein (Felsenstein, 1981):

$$L_i(\theta) = \sum_{a=0}^{\max_{int}} \pi_a \Pr^{(ROOT)}(a) \quad (3)$$

where max_{int} is the maximum number of family members in internal phylogenetic nodes (bounded for computational reasons), π_a is the prior probability of the number of members at the root, and $\Pr^{(ROOT)}(a)$ is the probability of having 'a' family members in the most internal node (the phylogenetic root).

We defined the max_{int} value as:

$$\max_{\text{int}} = 2\max_{ext} + n \quad (4)$$

where max_{ext} is the maximum number of members in external nodes, and n is an user-defined constant value (by default, $n = 10$). The rationale behind this dynamic upper boundary (for each particular gene family) is to optimize the trade-off between the computational time and the numerical precision of the likelihood computation.

BadiRate models π_a by using a Poisson or a Negative Binomial distribution, with parameters estimated by likelihood (ML or MAP) or by parsimony methods. For the latter case, BadiRate assumes that the parameter of the Poisson distribution is equal to the family size inferred by parsimony.

The pruning algorithm of Felsenstein assumes that lineages evolve independently of each other, so BadiRate computes $\Pr^{(ROOT)}(a)$ as the product of the probabilities of the two descendant lineages. For instance, the probability to have a family of '*a'* members in the most internal node, is given by the following recursive equation:

$$\Pr^{(ROOT)}(a) = \left(\sum_{d_y=0}^{\max_{int}} \Pr(a \mid d_y, t_y, \theta_y) \Pr^{(y)}(d_y) \right) \left(\sum_{d_z=0}^{\max_{int}} \Pr(a \mid d_z, t_z, \theta_z) \Pr^{(z)}(d_z) \right) \qquad (5)$$

where $Pr^{(y)}(d_y)$ is the probability of having d_y members in the descendant node y, and $\Pr(a|d_y,t_y,\theta_y)$ is the transition probability from a state 'a' (number of members in the internal node) to a state d_y. We calculated the transition probabilities of the stochastic GD (Gain-and-Death) and BDI (Birth-Death-and-Innovation, also known as Birth-Death-and-Immigration) models as in (Csuros and Miklos, 2009). In addition, BadiRate can also model the process assuming that birth and death rates are equal, a specific BDI case (nearly equivalent to the model implemented in CAFE) denoted here as the lambda-innovation (LI) model.

BadiRate can estimate the ML family turnover rates by assuming equal rates across lineages (global rate model), by considering variation in specific branches (branch-specific rate models), or even by assuming independent turnover rates in all lineages (free rate model). ML estimates under different branch models can be further used (for instance, by means of the likelihood ratio test, Akaike, or Bayesian information criteria) to test relevant biological hypothesis (*e.g.,* the

existence of higher or lower family turnover rates in specific lineages or groups of lineages). BadiRate allows detecting outlier lineages but also reports the gene families that have not likely evolved under the ML turnover rates (*i.e.*, families that significantly depart from the ML estimates); these outliers are interesting candidates for lineage-specific adaptations. Moreover, BadiRate uses these ML estimates to infer the most likely number of family members in the internal nodes, and it also determines the minimum number of gains and losses in each lineage.

Maximum a Posteriori (MAP) estimates

BadiRate can also incorporate prior biological information about family turnover rates. For instance, we can model the fact that HGT or *de-novo* origin of genes are unlikely events (Zhou, et al., 2008), or the massive gene loss that usually occur in endosymbiotic lineages (Mira, et al., 2001), by defining appropriate *a priori* probability distributions for the innovation and death rates, respectively. Specifically the MAP is given by:

$$\theta_{MAP} = \arg\max_{\theta} L(\theta) g(\theta) \qquad (6)$$

where $g(\theta)$ is the *a priori* distribution of the family turnover rates. BadiRate models $g(\theta)$ with a Gamma distribution with the scale parameter fixed at the value 1 and, therefore, the mean and variance are uniquely determined by the shape (α) parameter; thus, we may specify small α values if the biological information are indicative of low turnover rates, and *vice versa*. For instance, to analyze eukaryotic multigene families, we might use a Gamma distribution with a low shape value *(e.g.* $\alpha = 0.00001$, or even $\alpha = 0$) as the innovation rate prior; conversely, to estimate family turnover rates in endosymbiotic lineages, we might assume a Gamma distribution with a high shape value (*e.g.* $\alpha = 0.5$) as the death rate prior.

Parsimony estimates

In addition to the BDI, LI and GD stochastic models, we have also implemented an algorithm to estimate the family turnover rates by parsimony. BadiRate first infers the number of genes in all internal nodes by using the Wagner parsimony or a modification of the Sankoff algorithm. Then, BadiRate uses this information to determine the minimum number of gain and loss events in each branch; assuming one innovation event only if there are zero gene copies in a given internal node and one (or more) in the descendent nodes (otherwise, the family gains will be considered gene duplications). From this number of gain and loss events, BadiRate finally calculates the density-dependent (r_d, for birth and death) and independent (r_i, for gain and innovation) family turnover rates as follows (Vieira, et al., 2007):

$$r_d = \sum_{b=1}^{n} (e_b / a_b) / t \qquad (7)$$

$$r_i = \frac{\sum_{b=1}^{n} e_b}{t} \qquad (8)$$

where n is the number of phylogenetic lineages, t the total divergence time of the phylogeny (the sum of all branch lengths), and a_b and e_b are the number of ancestral family members and events (births, deaths, gains or innovations) in lineage b, respectively.

The accuracy of the parsimony algorithm relies on its ability to correctly infer evolutionary events. Because the algorithm attempts to minimize the number of events, the gene turnover rates may be underestimated. To avoid a situation where gain and loss events mask each other, it is helpful to separate a particular gene family into its component ortholog groups (or subfamilies); the inference of gains and losses separately in each subfamily leads to better parsimony estimates. It is worth noting that BadiRate can also use the parsimony inferences as the starting values to reduce convergence problems in the ML estimates of the family turnover rates.

Error Estimates (Computer Simulations)

To assess the performance of BadiRate on simulated data, we measured the error between estimated and true (fixed in simulations) turnover rates as the normalized Euclidean distance (*d*) between each family in each replicate and the simulated fixed value:

$$d(e_i,t) = \sqrt{\left(\frac{\left(\beta_{ei} - \beta_t\right)^2}{{\sigma_{\beta e}}^2} + \frac{\left(\delta_{ei} - \delta_t\right)^2}{{\sigma_{\delta e}}^2} \right)} \qquad (9)$$

where β_t and δ_t are the true birth and death rates, β_{ei} and δ_{ei} those rates estimated in the *i* replicate, and $\sigma_{\delta e}$ and $\sigma_{\beta e}$ the standard deviation over all replicates. This distance has the advantage of taking into account standard deviations, which allows testing if the estimates significantly depart from the rates fixed in the simulations.

Results

Simulation Results

We assessed the accuracy of the family turnover rates estimates by conducting computer simulations on the well characterized 12 Drosophila species phylogeny (*Drosophila* 12 Genomes Consortium, 2007) (Supplementary Figures S1, S2, S3, S4 and S5). In the simulations, we fixed the number of genes in the root (S = 5, 10 and 20), applied a range of BDI or GD rates (λ, β and δ = 0, 0.002, 0.004, 0.006, 0.008 and 0.01; ι = 0, 0.001 and 0.01), and determined the actual number of genes at the leaves. For a particular scenario, we performed 100 replicates, each comprising 30 gene families (a total of 3000 trees). We then benchmarked the BadiRate ML (under the BDI, LI and GD models) and parsimony estimates, as well as the CAFE ML estimates (under the BDM model). Moreover, we also evaluated the performance of the BadiRate ML estimates under different root family size distributions: under a Poisson with the parameter estimated by ML (P_{ML}) or parsimony (P_P), and under a Negative Binomial with parameters estimated by ML (NB_{ML}).

Our results show that, in general (and as expected), the family turnover rates estimated by parsimony are less accurate than the ML estimates (Figure 1A). Among the ML turnover estimates, the BDI method implemented in BadiRate outperforms the two others (LI and CAFE); this is especially true in cases where small-size families have asymmetric birth/death rates, i.e. $\beta > \delta$ or vice versa (Figure 1A and 1B; Supplementary Figure S2). Apart from the higher accuracy of the BDI model, which mainly results from the separate estimation of birth and death rates, the LI model also outperforms CAFE, even in scenarios with a null innovation rate (Figures 1A and 1C; Supplementary Figure S3). Unlike LI and CAFE, which are particular cases of the BDI model, the performance of the GD model is not directly comparable to BDI. Still, our simulations show good performance of the GD model in the analyzed scenarios (Supplementary Figure S4).

We also found that the BadiRate ML estimates are more accurate using the P_{ML} or NB_{ML} (root family size distributions) in cases of asymmetric birth/death rates, and using the P_P in cases of similar birth and death rates. This result might reflect the fact that our parsimony algorithm equally scores gain and loss events, which yields good root family size inferences only in cases of similar birth and death rates. Its performance and favorable running speed, makes the P_P approach a good alternative for the analysis of families with similar birth and death rates (Supplementary Figure S5).

Empirical Data Results

We also illustrate a BadiRate application by analyzing published microRNA data from the 12 Drosophila species (Nozawa, et al., 2010) under the GD model. In this work, the authors suggested a putative miRNA expansion in the *D. willistoni* lineage. To test this hypothesis, we compared the likelihood of the turnover rates under a global-rates model to a model assuming specific turnover rates in the *D. willistoni* branch. The likelihood-ratio test is not significant (p = 0.1527), suggesting that there are no differences on the miRNA turnover rates between *D.*

willistoni and the rest of *Drosophila* lineages. However, because miRNAs in the 11 non-melanogaster species were identified by similarity based on the available *D. melanogaster* miRNA data, the identification of miRNA family members is less accurate for longer divergence times. This can lead to the detection of spurious miRNA expansions, which may mask putative changes in the turnover rates in *D. willistoni*. To control for this effect, we compared the likelihood values of two scenarios. The first scenario assumes independent turnover rates in two classes of branches (in the internal lineages leading to *D. melanogaster* and in the rest of branches), while the second additionally incorporates a third class of turnover rates (for the *D. willistoni* lineage). The lower Akaike Information Criterion (AIC) value of the second scenario (AIC = 1509.8128) compared to the first one (AIC = 1518.3918), and the significant LRT (P = 0.0018), indicates that the *D. willistoni* lineage indeed has distinct miRNA turnover rates. Moreover, we also inferred the most likely number of miRNA elements in the internal nodes of the *Drosophila* phylogeny; these figures are very similar to that estimated by (Nozawa, et al., 2010) (Supplementary Figure S6).

BadiRate's Input and Output

BadiRate requires as input the established species phylogenetic tree (ultrametric-rooted tree in Newick format) and a tab-delimited data file with the family size for each species represented in the phylogeny (for full details see BadiRate's documentation). The output includes estimates of the turnover rates, the most likely number of members in internal nodes (in extended Newick format), and the outlier families (families with turnover rates significantly higher or lower than that estimated for the whole data after correcting for multiple testing).

REFERENCES

1. Cohen, O., Rubinstein, N.D., Stern, A., Gophna, U. and Pupko, T. (2008) A likelihood framework to analyse phyletic patterns, *Philos Trans R Soc Lond B Biol Sci*, **363**, 3903-3911.

2. Csuros, M. and Miklos, I. (2009) Streamlining and large ancestral genomes in Archaea inferred with a phylogenetic birth-and-death model, *Mol Biol Evol*, **26**, 2087-2095.

3. *Drosophila* 12 Genomes Consortium (2007) Evolution of genes and genomes on the Drosophila phylogeny, *Nature*, **450**, 203-218.

4. Felsenstein, J. (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach, *J Mol Evol*, **17**, 368-376.

5. Felsenstein, J. (1992) Phylogenies from restriction sites: A maximum-likelihood approach, *Evolution*, **46**, 16.

6. Mira, A., Ochman, H. and Moran, N.A. (2001) Deletional bias and the evolution of bacterial genomes, *Trends Genet*, **17**, 589-596.

7. Nelder, J.A. and Mead, R. (1965) A Simplex Method for Function Minimization, *The Computer Journal*, **7**, 308-313.

8. Nozawa, M., Miura, S. and Nei, M. (2010) Origins and evolution of microRNA genes in Drosophila species, *Genome Biol Evol*, **2**, 180-189.

9. Vieira, F.G., Sanchez-Gracia, A. and Rozas, J. (2007) Comparative genomic analysis of the odorant-binding protein family in 12 Drosophila genomes: purifying selection and birth-and-death evolution, *Genome Biol*, **8**, R235.

10. Zhou, Q., Zhang, G., Zhang, Y., Xu, S., Zhao, R., Zhan, Z., Li, X., Ding, Y., Yang, S. and Wang, W. (2008) On the origin of new genes in Drosophila, *Genome Res*, **18**, 1446-1455.

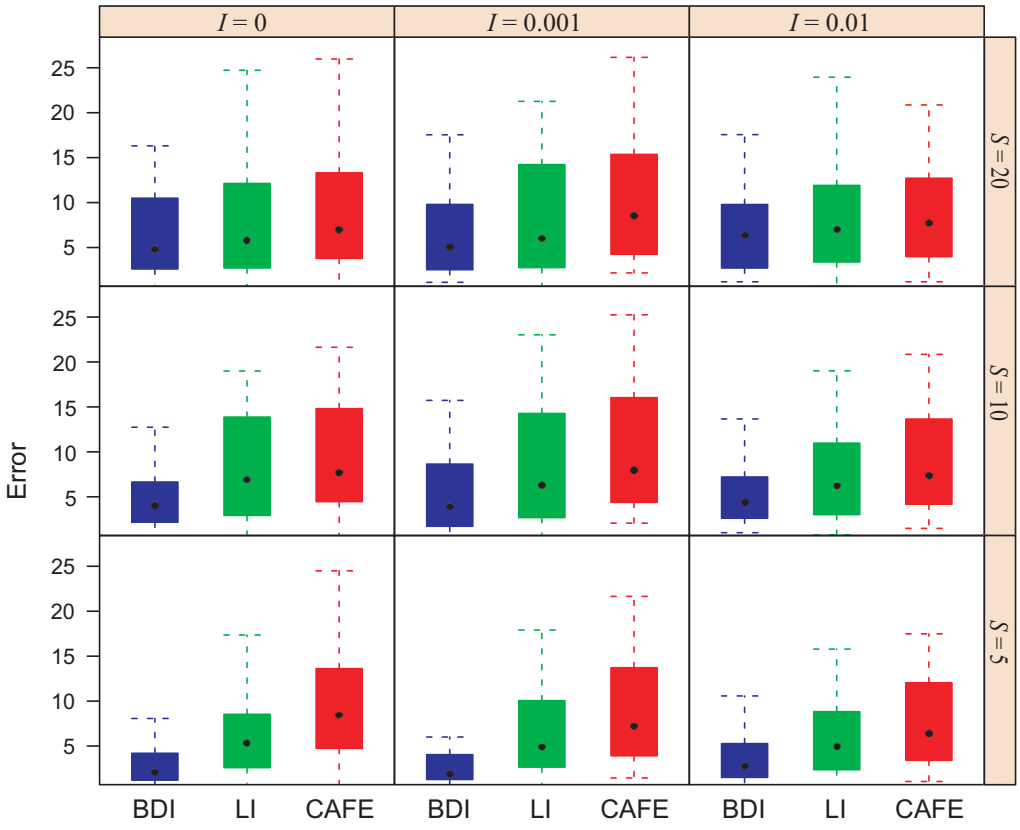


Figure S1 - Relative performance of the different methods. Gene family simulation with different number of members at the root ($S = 5$, 10 and 20) and innovation rates ($I = 0$, 0.001 and 0.01). The boxplots represent the distribution of error values across the tested scenarios. The error of the ML estimates under the BDI and LI models (conducted with BadiRate) are depicted in blue and green, respectively. The error of ML estimates conducted with CAFE (BDM model) is depicted in red.

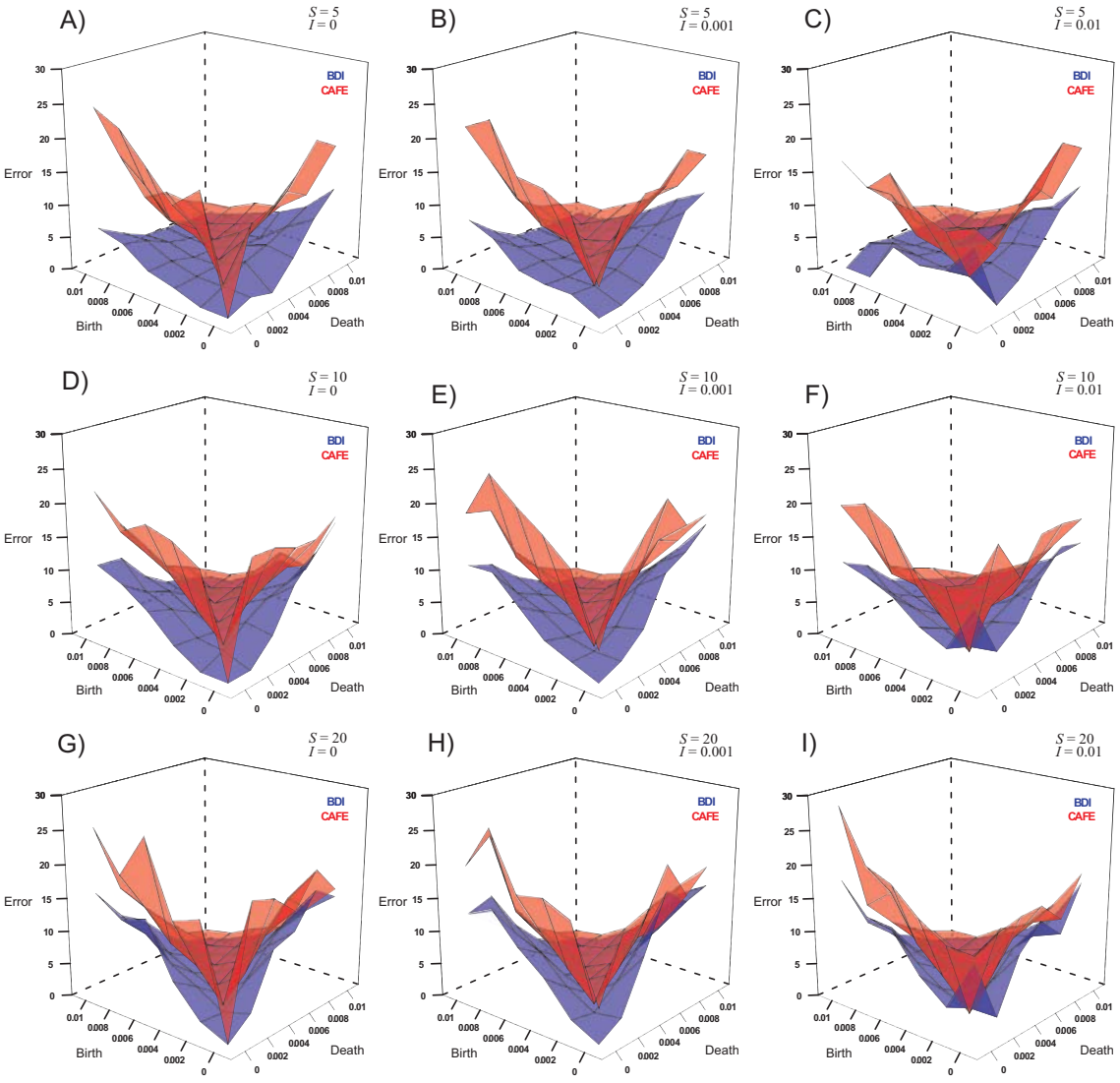


Figure S2 - Comparision of the performance of the ML estimates obtained by BadiRate and CAFE. The errors of the ML estimates are calculated as the normalized Euclidean distance between the true value (fixed in the simulations) and that obtained by BadiRate (BDI model) and CAFE (BDM model). Surface plots represent the error values for different birth and death rate scenarios. The error values of BadiRate (BDI model) and CAFE (BDM model) are depicted in blue and red, respectively. *S*, number of genes at the most internal node; *I*, innovation rate

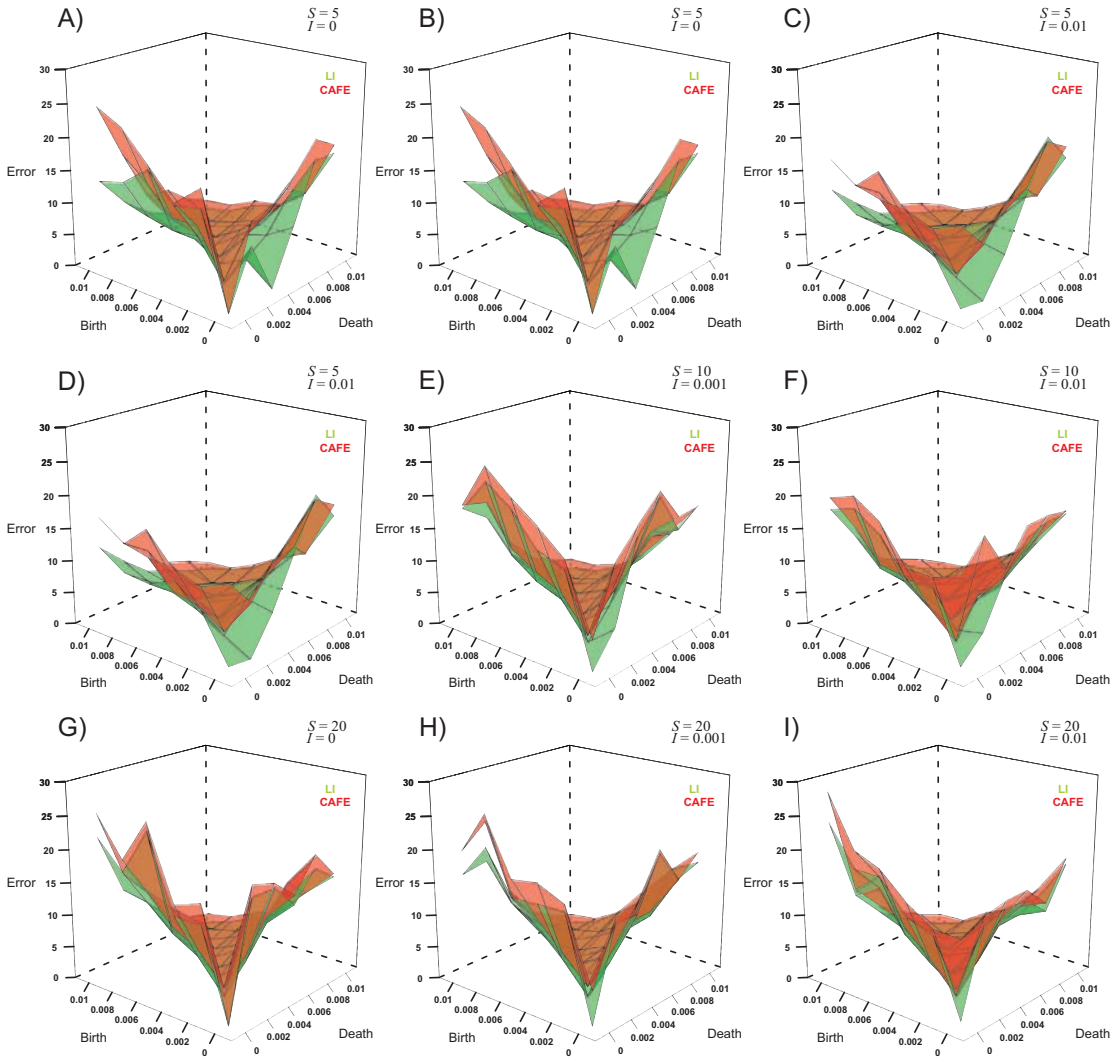


Figure S3 - Comparision of the performance of the ML estimates obtained by BadiRate and CAFE. The errors of the ML estimates are calculated as the normalized Euclidean distance between the true value (fixed in the simulations) and that obtained by BadiRate (LI model) and CAFE (BDM model). Surface plots represent the error values for different birth and death rate scenarios. The error estimates of BadiRate (LI model) and CAFE (BDM model) are depicted in green and red, respectively. *S*, number of genes at the most internal node; *I*, innovation rate.

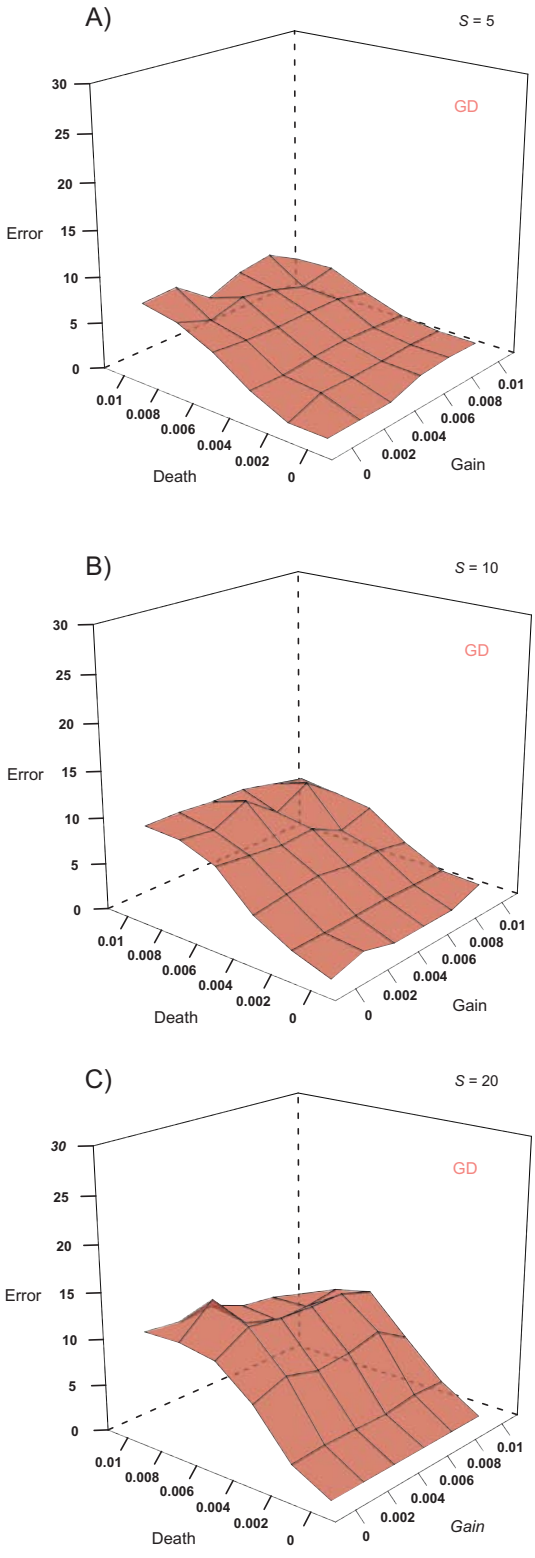


Figure S4 - Performance of BadiRate under the GD model. The errors of the ML estimates are calculated as the normalized Euclidean distance between the true value (fixed in the simulations) and that obtained by BadiRate (GD model). Surface plots represent the error values for different gain and death rate scenarios. *S*, number of genes at the most internal node

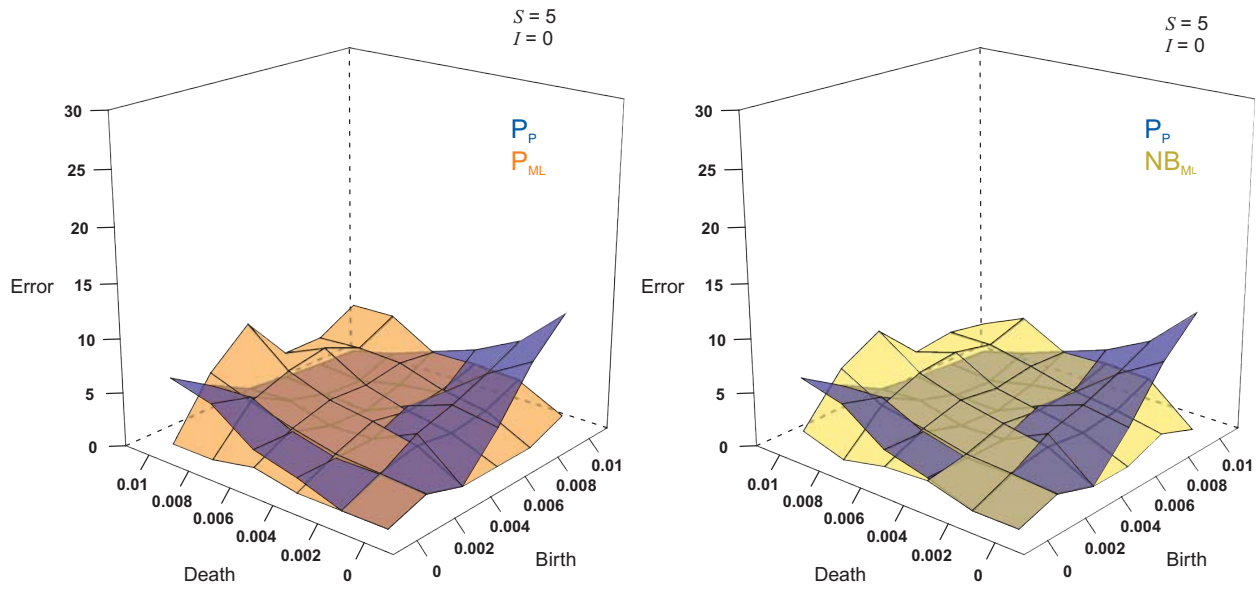


Figure S5 - Comparision of the performance of different *a priori* root family size distributions. The errors of the ML estimates are calculated as the normalized Euclidean distance between the true value (fixed in the simulations) and that obtained by BadiRate (BDI model). The *a priori* root family size distributions P_P, P_{ML} and a NB_{ML} are depicted in blue, orange and yellow, respectively. *S*, the number of genes at the most internal node; *I*, innovation rate.

3.3 PopDrowser: the Population Drosophila Browser

La secuenciación de 168 líneas isogénicas de una única población natural (Raleigh, Carolina del Norte) de *D. melanogaster* es un recurso sin precedentes para la comunidad científica, ya que posibilita detectar procesos incipientes de adaptación.
En este artículo se presenta el '*Population Drosophila Browser*' (PopDrowser), un nuevo navegador genómico especialmente diseñado para analizar y visualizar la variación nucleotídica a lo largo del genoma de *D. melanogaster*. PopDrowser incluye una serie de estadísticos descriptivos del nivel y patrón de polimorfismo y divergencia nucleotídico, estimas del desequilibrio de ligamiento y varios *tests* de neutralismo. Además, PopDrowser puede re-calcular los estadísticos con parámetros definidos por el usuario (ej. longitud de la región analizada, excluir *singletons*, etc.), lo que le dota de una gran versatilidad para posteriores estudios evolutivos.

BIOINFORMATICS APPLICATIONS NOTE

Vol. 28 no. 4 2012, pages 595–596
doi:10.1093/bioinformatics/btr691

Genetics and population analysis

Advance Access publication December 15, 2011

PopDrowser: the Population Drosophila Browser

Miquel Ràmia[1,†], Pablo Librado[2,†], Sònia Casillas[1,†], Julio Rozas[2] and Antonio Barbadilla[1,*]

[1]Institut de Biotecnologia i de Biomedicina and Departament de Genètica i de Microbiologia (Facultat de Biociències), Universitat Autònoma de Barcelona, 08193 Bellaterra (Barcelona) and [2]Departament de Genètica and Institut de Recerca de la Biodiversitat (IRBio), Universitat de Barcelona, Diagonal 645, 08028 Barcelona, Spain

Associate Editor: Jeffrey Barret

ABSTRACT

Motivation: The completion of 168 genome sequences from a single population of *Drosophila melanogaster* provides a global view of genomic variation and an understanding of the evolutionary forces shaping the patterns of DNA polymorphism and divergence along the genome.

Results: We present the 'Population Drosophila Browser' (PopDrowser), a new genome browser specially designed for the automatic analysis and representation of genetic variation across the *D. melanogaster* genome sequence. PopDrowser allows estimating and visualizing the values of a number of DNA polymorphism and divergence summary statistics, linkage disequilibrium parameters and several neutrality tests. PopDrowser also allows performing custom analyses on-the-fly using user-selected parameters.

Availability: PopDrowser is freely available from http://PopDrowser.uab.cat.

Contact: miquel.ramia@uab.cat

Received on October 6, 2011; revised on November 29, 2011; accepted on December 8, 2011

1 INTRODUCTION

Population genetics studies have been so far based on fragmentary and non-random samples of genomes, providing a partial and often biased view of the population genetics processes (Begun *et al*., 2007). A new dimension to genetic variation studies is provided by the new availability of within-species genomes. Next-generation sequencing technologies are making affordable genome-wide population genetics data, not only for humans and the main model organisms, but also for most organisms on which research is actively carried out on genetics, ecology or evolution (Pool *et al.*, 2010).

Genome browsers are very useful tools to query and visualize disparate annotations at different genomic locations using a web user interface (Schattner, 2008). A number of web-based genome browsers displaying genetic variation data are already available (Benson *et al.,* 2002; Dubchak and Ryaboy, 2006; Frazer *et al.*, 2007; Hubbard *et al*., 2002; Kent *et al.*, 2002; Stein *et al.*, 2002). Such browsers, however, are not well suited to deal with population genomics sequence information. For example, HapMap (International HapMap Consortium, 2003), the most comprehensive genome browser of variation data so far, contains information on single nucleotide polymorphisms (SNPs), Copy Number Variations (CNVs) and linkage disequilibrium of human populations. It does not offer, however, genetic variation estimates along sliding-windows or neutrality-based tests.

The Drosophila Genetic Reference Panel (DGRP) (T.Mackay *et al.*, accepted for publication) has recently sequenced and analyzed the patterns of genome variation in 168 inbred lines of *Drosophila melanogaster* from a single population of Raleigh (USA), and conducted a genome-wide association analysis of some phenotypic traits. A major goal of this project is to create a resource of common genetic polymorphism data to aid further population genomics analyses. As a part of this DGRP project, here we present a modified Gbrowse specifically designed for the automatic estimation and representation of population genetic variation in *D. melanogaster*, the 'Population Drosophila Browser' (PopDrowser). Unlike other population analysis tools (Hutter *et al.*, 2006; Kofler *et al.*, 2011), the PopDrowser is a genome browser, which can be customized to create analogous resources for any other species with within-species polymorphism data.

2 IMPLEMENTATION

2.1 Input data

The initial input data are a set of 168 aligned intraspecific *D. melanogaster* sequences from the DGRP project, and also include the genome sequences of *Drosophila yakuba* and *Drosophila simulans,* which were used as outgroup species.

2.2 Interface and implementation

PopDrowser allows reporting precomputed estimates of several DNA variation measures along each chromosome arm through the combined implementation of the programs PDA 2 (Casillas and Barbadilla, 2006), MKT (Egea *et al.*, 2008) and VariScan 2 (Hutter *et al.*, 2006). The data and summary statistics are graphically displayed along the chromosome arms on a web-based user interface using the Gbrowse software.

PopDrowser also includes an innovative capability that allows performing custom analyses on-the-fly. After selecting a chromosome region and a particular track, the user can conduct exhaustive analyses by defining their own custom input parameters. Furthermore, users can choose to either visualize the output of their analyses graphically in the browser—as a new track—or to

*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the first three authors should be regarded as joint First Authors.

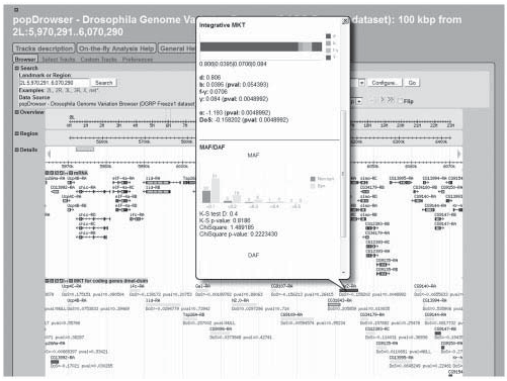


Fig. 1. PopDrowser snapshot showing the results of the McDonald–Kreitman tests in the *ade2-RA* gene within its genome context.

download it in a tabulated text file. The estimates available for on-the-fly analyses are specified in Table1.

The current implementation is running in an Ubuntu 10.04 Linux x64 server, 2 IntelXeon 3Ghz processors, 32GB RAM with Apache.

2.3 Output

Along with reference genome annotations, the genome browser output includes measures of a number of nucleotide summary statistics, such as the levels of nucleotide diversity (π and θ), DNA divergence between species (K), different measures of linkage disequilibrium and genome-wide neutrality tests. Such analyses are computed along each chromosome arm in non-overlapping sliding windows of 0.05, 0.1, 0.5, 1, 10, 50 and 100 kb. For each gene, the browser also provides a single track including information of the generalized and the integrative McDonald–Kreitman tests (McDonald and Kreitman, 1991; T.Mackay *et al.*, accepted for publication) along with minor and derived allele frequency (MAF, DAF) spectrums (Fig. 1). All the tracks included in the PopDrowser are summarized in Table 1.

Table 1. Summary of the PopDrowser tracks

Category	Gene annotations and estimates
Drosophila melanogaster reference annotations (build 5.13) and recombination	Gene structure, mRNA, CDS (Coding Sequence), ncRNA, tRNA, orthologous genes, phastCons, GC content. local recombination rate (Fiston-Lavier *et al.*, 2010)
Density tracks	Genes, microsatellites, transposons, CDS, SNPs
Nucleotide variants	SNPs, single nucleotide fixations
Measures of nucleotide variation and LD)[a]	Number of segregating sites (S), total minimum number of mutations (η), number of singletons (η_e), nucleotide diversity (π), Watterson's estimator of nucleotide diversity per site (θ), number of haplotypes (h), haplotype diversity (Hd), nucleotide divergence per site (corrected by Jukes–Cantor) (K). LD: D, absolute D ($\|D\|$), D', absolute D' ($\|D'\|$), r^2
Neutrality tests[a]	Fu and Li's D, D^a, F, F^a, Fay and Wu's H, Tajima's D, Fu's F_S statistics. MKT (per gene)

LD, linkage disequilibrium; CDS, coding sequence.
[a]Estimates available for on-the-fly analyses (except MKT per gene).

ACKNOWLEDGEMENTS

We thank Raquel Egea for helping in implementing the MKT software and Albert Vernon Smith from HapMap for his help in the implementation of the pie graph glyph. We also thank Dave Clements, Lincoln Stein and Scott Cain for technical support, and John H. Werren for valuable discussions on the browser. This paper was prepared with full knowledge and support of the DGRP Consortium.

Funding: Ministerio de Ciencia e Innovación (Spain) (BFU2009-09504 to A.B., BFU2010-15484 to J.R.); Catalonian Comissió Interdepartamental de Recerca i Innovació Tecnològica (2009SGR-88 to A. Ruiz; 2009SGR-1287 to M. Aguadé); Departament de Genètica i de Microbiologia of the Universitat Autònoma de Barcelona (409-04-2/08 to M.R.); ICREA Academia (Generalitat de Catalunya to J.R., in part).

Conflict of Interest: none declared.

REFERENCES

Begun,D.J. *et al.* (2007) Population genomics: whole-genome analysis of polymorphism and divergence in Drosophila simulans. *PLoS Biol.*, **5**, e310.

Benson,D.A. *et al.* (2002) GenBank. *Nucleic Acids Res.*, **30**, 17–20.

Casillas,S. and Barbadilla,A. (2006) PDA v.2: improving the exploration and estimation of nucleotide polymorphism in large datasets of heterogeneous DNA. *Nucleic Acids Res.*, **34**, W632–W634.

Dubchak,I. and Ryaboy,D.V. (2006) VISTA family of computational tools for comparative analysis of DNA sequences and whole genomes. *Methods Mol. Biol.* 2006, **338**, 69–89.

Egea,R. *et al.* (2008) Standard and generalized McDonald-Kreitman test: a website to detect selection by comparing different classes of DNA sites. *Nucleic Acids Res.*, **36**, W157–W162.

Fiston-Lavier,A.S. *et al.* (2010) *Drosophila melanogaster* recombination rate calculator. *Gene*, **463**, 18–20.

Frazer,K.A. *et al.* (2007) A second generation human haplotype map of over 3.1 million SNPs. *Nature*, **449**, 851–861.

Hubbard,T. *et al.* (2002) The Ensembl genome database project. *Nucleic Acids Res.*, **30**, 38–41.

Hutter,S. *et al.* (2006) Genome-wide DNA polymorphism analyses using VariScan. *BMC Bioinformatics*, **7**, 409.

International HapMap Consortium (2003) The International HapMap Project. *Nature*, **426**, 789–796.

Kent, W.J. *et al.* (2002) The human genome browser at UCSC. *Genome Res.*, **12**, 996–1006.

Kofler,R. *et al.* (2011) PoPoolation, a toolbox for population genetic analysis of next generation sequencing data from pooled individuals. *PLoS One*, **6**, e15925.

McDonald,J.H. and Kreitman,M. (1991) Adaptive protein evolution at the Adh locus in Drosophila. *Nature*, **351**, 652–654.

Pool,J.E. *et al.* (2010) Population genetic inference from genomic sequence variation. *Genome Research*, **20**, 291–300.

Schattner,P. (2008) *Genomes, Browsers and Databases. Data-Mining Tools for Integrated Genomic Databases*. Cambridge University Press, New York.

Stein,L.D. *et al.* (2002) The generic genome browser: a building block for a model organism system database. *Genome Res.*, **12**, 1599–1610.

3.4 Uncovering the functional constraints underlying the genomic organization of the odorant-binding protein genes

El sistema olfativo tiene un papel fundamental en la supervivencia y reproducción de los individuos. En insectos, las primeras etapas de la quimiopercepción están mediadas por las proteínas de unión a odorantes (OBPs). La distribución de los genes que codifican para OBPs no es aleatoria, sino que se conserva más tiempo de lo esperado dadas las tasas de reordenamiento cromosómicas observadas en *Drosophila*. Hasta la fecha, todavía no se conocían las causas de dicha conservación.

En este estudio presentamos un análisis exhaustivo de los factores potencialmente responsables de dicha conservación, que incluyen: i) la arquitectura de la región promotora, ii) un ambiente transcripcional característico, y iii) el estado de la cromatina.

Nuestros resultados sugieren que los dominios de la cromatina restringen la ubicación de los genes que codifican para OBPs en regiones cromosómicas que se caracterizan por un determinado ambiente transcripcional. Sin embargo, y de forma aparentemente contradictoria con los modelos establecidos, este ambiente no se caracteriza por un menor ruido transcripcional (EN). De hecho, el EN podría incrementar la plasticidad fenotípica de los individuos, lo cual puede ser crítico en la percipción olfativa.

GBE

Uncovering the Functional Constraints Underlying the Genomic Organization of the Odorant-Binding Protein Genes

Pablo Librado and Julio Rozas*

Departament de Genètica and Institut de Recerca de la Biodiversitat (IRBio), Universitat de Barcelona, Barcelona, Spain

*Corresponding author: E-mail: jrozas@ub.edu.

Accepted: October 17, 2013

Abstract

Animal olfactory systems have a critical role for the survival and reproduction of individuals. In insects, the odorant-binding proteins (OBPs) are encoded by a moderately sized gene family, and mediate the first steps of the olfactory processing. Most OBPs are organized in clusters of a few paralogs, which are conserved over time. Currently, the biological mechanism explaining the close physical proximity among OBPs is not yet established. Here, we conducted a comprehensive study aiming to gain insights into the mechanisms underlying the OBP genomic organization. We found that the OBP clusters are embedded within large conserved arrangements. These organizations also include other non-OBP genes, which often encode proteins integral to plasma membrane. Moreover, the conservation degree of such large clusters is related to the following: 1) the promoter architecture of the confined genes, 2) a characteristic transcriptional environment, and 3) the chromatin conformation of the chromosomal region. Our results suggest that chromatin domains may restrict the location of OBP genes to regions having the appropriate transcriptional environment, leading to the OBP cluster structure. However, the appropriate transcriptional environment for OBP and the other neighbor genes is not dominated by reduced levels of expression noise. Indeed, the stochastic fluctuations in the OBP transcript abundance may have a critical role in the combinatorial nature of the olfactory coding process.

Key words: chemosensory system, olfactory reception, gene cluster constraint, expression noise, chromatin domain.

Introduction

Animal olfactory systems allow for the detection of food, predators, and mates, and thus demonstrating a critical role for the survival and reproduction of individuals (Krieger and Ross 2002; Matsuo et al. 2007). In *Drosophila*, the early steps of odor processing occur in chemosensory hairs (i.e., the sensilla), which are located in the third antennal segment and the maxillary palp. The main biochemical events include the uptake of volatile molecules through the cuticle pores, transport across the sensilla lymph, and interaction with olfactory receptors. The latter steps are mediated by the odorant-binding proteins (OBPs), which may have an active role in olfactory coding such as contributing to odor discrimination (Swarup et al. 2011) and receptor activation (Laughlin et al. 2008; Biessmann et al. 2010). OBPs are small (10–30 kDa; 130–220 aa long), highly abundant, globular, and water-soluble proteins (Kruse et al. 2003; Tegoni et al. 2004). These molecules are encoded by a moderately sized multigene family (in the 12 *Drosophila* species, the number of OBP members range from 41 to 62), with an evolution that is consistent with the birth-and-death model (Vieira et al. 2007).

In arthropods, most OBP genes are organized in clusters of a few paralogs (Hekmat-Scafe et al. 2002; Foret and Maleszka 2006), an arrangement that is moreover conserved over time (Vieira and Rozas 2011). Nevertheless, it is not well established whether the conservation of these OBP clusters represent the outcome of an uneven distribution of chromosomal rearrangement breakpoints, or rather they are constrained by natural selection for some functional meaning (Zhou et al. 2009; Sanchez-Gracia and Rozas 2011; Vieira and Rozas 2011). For example, functionally linked genes, such as those encoding subunits of the same complex (Chamaon et al. 2002), proteins of the same pathway (Lee and Sonnhammer 2003), or genes with expression patterns restricted to the head, embryo, or testes (Boutanaev et al. 2002) are often clustered in the *Drosophila melanogaster* genome. As clusters of functionally linked genes may include nonhomologous members, the OBP gene organization may be preserved by functional constraints imposed from neighboring genes.

The presence of shared cis-regulatory elements, such as bidirectional promoters or pleiotropic enhancers, may explain the OBP gene organization (Li et al. 2006; Yang and Yu 2009).

For example, central regions of some *Drosophila* syntenic clusters are enriched for highly conserved noncoding elements that regulate the transcription of genes with the appropriate composition of core promoter elements (CPEs) (Engstrom et al. 2007). Notably, the CPE composition and expression pattern are two features characterizing the broad and peaked promoter architectures (a classification based on the distribution of transcription start sites) (Hoskins et al. 2011). Although genes with peaked promoters are often expressed in specific tissues or developmental stages, those with broad promoters usually have constitutive transcription (Kharchenko et al. 2011; Rach et al. 2011). Therefore, shared cis-regulatory elements may differentially restrict the movement of genes with particular promoter architectures or transcriptional patterns.

Chromatin conformation (Filion et al. 2010; Kharchenko et al. 2011) could also affect gene organization given its role in the regulation of gene expression (i.e., the so-called position effect). For example, human unfolded chromatin (30-nm chromatin fibers) encompasses high-density gene regions (Gilbert et al. 2004), which usually exhibit elevated expression breadth (EB) (Caron et al. 2001; Lercher et al. 2002). Interestingly, transcriptional activation after chromatin unfolding induces stochastic fluctuations in transcript abundance (i.e., expression noise [EN]) (Becskei et al. 2005). Such EN is often deleterious, particularly for broadly expressed genes, because it yields imbalances in the stoichiometry of proteins (Fraser et al. 2004). These features led Batada and Hurst (2007) to hypothesize that broadly expressed genes are clustered in regions of constitutively unfolded chromatin to minimize EN. Several lines of thought support this model. For example, as *head-to-head* gene pairs share their promoter regions, a chromatin unfolding event can facilitate the transcriptional activation of both genes. Therefore, chromatin unfolding events will be less frequent in head-to-head than other gene pair arrangements, leading to reduced levels of EN (Wang et al. 2010). Because EN is often deleterious, natural selection may favor the maintenance of the head-to-head gene pair organization in clusters.

Chromosomal proteins determining the chromatin state, such as nuclear membrane (Capelson et al. 2010; Vaquerizas et al. 2010), insulators (Maeda and Karch 2007; Wallace et al. 2009; Negre et al. 2010), and chromatin remodeling (Kalmykova et al. 2005; Li and Reinberg 2011) proteins, may therefore play a relevant role in maintaining gene clusters. In this regard, the function of the JIL-1 protein kinase deserves special attention for its role in defining the decondensed interbands of polytene chromosomes, which characterize active and unfolded chromatin (Jin et al. 1999; Regnard et al. 2011; Kellner et al. 2012). Moreover, JIL-1 kinase, which phosphorylates Serine 10 and 28 at Histone 3, physically interacts with the lamin Dm0 (a structural nuclear membrane protein) (Bao et al. 2005) and Chromator (localized in the spindle matrix of the nucleosketeleton) (Gan et al. 2011) proteins. Recently, the lamin Dm0 protein has been shown to colocalize with conserved microsynteny in *Drosophila* (Ranz et al. 2011), whereas Chromator changes the chromatin folding state (Rath et al. 2006). Therefore, high-order regulatory mechanisms involving chromatin conformation may underlie the conservation of some gene clusters.

Here, we analyzed the mechanisms underlying the OBP genomic organization. We found that the OBP clusters are embedded within large arrangements, which also include other non-OBP genes. The conservation degree of such large arrangements is moreover related to a number of functional and expression features, such as a transcriptional environment not dominated by reduced levels of EN. Indeed, the stochastic fluctuations in the OBP transcript abundance may have a critical role in the combinatorial nature of the olfactory coding process.

Materials and Methods

DNA Sequence Data and Assignment of Orthologous Groups

We downloaded the *D. melanogaster* gene and protein sequences and their orthologous relationships (release fb_2011_04) with the additional 11 *Drosophila* species (Drosophila 12 Genomes 2007) from FlyBase (release 5.40). The orthology data set contains predicted and curated pairwise relationships between the *Drosophila* species (i.e., one-to-one, one-to-many, and many-to-many relationships). We clustered these ortholog pairs into groups with multiple species using the Markov Clustering Algorithm software with default parameters (inflation = 2 and scheme = 7).

Gene Clustering

We define a conserved cluster as a group of neighbor genes maintained over time; this definition allowed us to study clusters of linked genes, regardless whether they are homologous. To infer such conserved gene clusters, we used the MCMuSeC software (Ling et al. 2009), which permits that clusters can undergo internal rearrangement events (Luc et al. 2003), as well as tandem gene duplications (recent duplicates originated from members of the same cluster). For each inferred cluster, we measured the conservation level as the branch length score (BLS), that is, the total divergence time (Tamura et al. 2004) since the cluster origin. The larger the BLS value, the more ancient the gene cluster.

We evaluated the significance of each BLS value separately for each cluster size (n). Indeed, small-sized clusters (with a low number of genes) have a lower probability to be disrupted by chromosomal rearrangements than larger ones. For each cluster size, we generated an empirical null distribution of the expected BLS value by randomly sampling 10,000 groups of n contiguous *D. melanogaster* genes, and the BLS values were computed across the information of the 12 *Drosophila* species. We defined the probability of an observed BLS value

(pBLS) as the fraction of sampled clusters with a BLS value lower than or equal to the observed (supplementary table S1, Supplementary Material online).

We also used computer simulations to examine whether the chromatin and expression factors that correlate with the pBLS value (e.g., JIL-1 binding intensity or EN) are specific constraints of the OBP gene organization, or correspond to genome-wide characteristics. We generated null empirical distributions by randomly sampling 10,000 replicates of 31 *D. melanogaster* clusters without OBP genes, but with the same number of genes and similar pBLS (±0.01) as that observed for clusters including OBP genes. For each replicate, we calculated the correlation between the characteristic chromatin and expression factors and the pBLS value. The probability of an observed correlation (*P* value) was estimated as the proportion of samples with correlation values higher than the observed. A low probability (i.e., $P < 0.05$) value indicates that the surveyed factor is not as common among the genome-wide *Drosophila* gene clusters as it is in the clusters including OBP genes.

Expression Data

We obtained gene expression data for all of the *D. melanogaster* genes from FlyAtlas (Chintapalli et al. 2007). We used the whole fly expression intensity (EI) information, and all of the 26 conditions incorporated in FlyAtlas, including larval and adult tissues. We considered that a gene is transcribed if the present call value was greater than zero. In addition to the EI value, we also computed the EB as the fraction of tissues where the gene is transcribed (regardless of the expression level in a given tissue), the sex-specific expression (SSE) as the transcription in sexual tissues (i.e., testis, ovary, male accessory glands, virgin spermateca, and mated spermateca) relative to the rest of tissues, and the EN as the coefficient of variation (COV) of the EI values. As the FlyAtlas expression data were determined from highly inbred flies (the Canton-S stock) reared at homogeneous conditions (22 °C with a 12 h:12 h light regime), the COV values are not explained by differences in the genetic or environmental background, but rather represent an excellent proxy to evaluate the stochastic fluctuations in transcript abundance (EN). The mean expression measures for each cluster were calculated as the average expression values of the spanned genes.

Functional Genomic Data

The ChIP-chip binding intensity for the JIL-1 protein and the nine chromatin states defined in Kharchenko et al. (2011) were downloaded from the modENCODE project database (BG3 *D. melanogaster* cell line). The nine-state chromatin model classifies each *D. melanogaster* nucleotide position into one out of nine chromatin states (i.e., Promoter and TSS, Transcription elongation [TE], Regulatory regions, Open chromatin, Active genes on the male X chromosome, Polycomb-mediated repression, Pericentromeric heterochromatin, Heterochromatin-like embedded in euchromatin, and Transcriptionally silent, intergenic) on the basis of the combinatorial profile of 18 histone marks (Kharchenko et al. 2011). The promoter architecture information, which integrates cap analysis of gene expression (CAGE), RNA ligase mediated rapid amplification of cDNA ends (RLM-RACE) and cap-trapped expressed sequence tags data, was obtained from Hoskins et al. (2011). We performed the promoter analysis using all promoter annotations, but also confirmed the results by restricting the analysis to promoters with only validated support (evidence from two or more data types; e.g., CAGE and RLM-RACE).

We used the FlyBase Gene Ontology (GO) annotation (release fb_2011_04) to gauge whether genes clustered with OBP genes are functionally related. We analyzed the GO over-representation using the Topology-Elim algorithm (Grossmann et al. 2007), which considers the hierarchical dependencies of the GO terms, and was implemented in the Ontologizer 2.0 software (Bauer et al. 2008).

Phylogeny-Based Analysis

The age of the genes (the divergence time since its origin) is a relevant factor to be considered when analyzing the mechanisms involved in gene cluster conservation. For example, recent gene duplications usually evolve faster than older ones (Luz et al. 2006) and often exhibit an SSE pattern. Moreover, the maximum BLS value of a particular cluster depends on the age of the encompassed genes. We inferred the maximum BLS cluster value as the minimum age of the encompassed genes, using the topological dating approach (Huerta-Cepas and Gabaldon 2011) with the BadiRate software (Librado et al. 2012).

Statistical Multivariate Analysis

We examined the relationships among the pBLS and a number of genomic and gene expression factors by different association tests (supplementary table S2, Supplementary Material online). On the one hand, we analyzed bivariate associations by using the following: 1) the Wilcoxon exact test, 2) the Pearson correlation coefficient, 3) the Spearman's rank correlation coefficient, and 4) the maximal information coefficient (MIC) (Reshef et al. 2011). We used the Wilcoxon exact test to compare clusters with low (<0.90) and high (>0.99) pBLS values. As this test requires a categorization of a continuous variable (the pBLS value), it is often conservative. For this reason, we also computed the Pearson correlation coefficient, which captures the linear continuous dependence between variables. Nevertheless, the Pearson correlation coefficient is very sensitive to outliers and skewed distributions, which may generate spurious associations between variables. Indeed, the assumptions required to calculate the probability associated to

the Pearson correlation coefficient may not hold in our data; for instance, the pBLS values are not normally distributed (Kolmogorov–Smirnov test: $P < 2.2e{-}16$). In such case, the Spearman's rank correlation coefficient is recommendable. This test, however, is not without problems, such as the use of the midrank approach for handling ties. The MIC-based test does not assume normality of the data and allows detecting a wide range of bivariate associations, including monotonic (e.g., linear, exponential) and nonmonotonic (e.g., sinusoidal) relationships. However, the *P* value of the MIC score can only be obtained by simulations. Currently only a few precomputed tables are available, which precludes computing exact *P* values, especially for our genome-wide data set (sample size of 3,434). Given these pros and cons, we reported the Spearman's rank correlation coefficient throughout the manuscript. In addition, it is worth noting that all conclusions extracted from the Spearman's rank correlation coefficient were also supported by other tests, especially the main findings (supplementary table S2, Supplementary Material online). On the other hand, as the examined variables are clearly intercorrelated, we also conducted a partial correlation and a path analysis. We assessed the goodness of fit of our empirical data to the underlying path model by evaluating the chi-squared significance.

The Wilcoxon exact test, the Pearson, and the Spearman's correlation coefficients, as well as the partial correlation and the path analysis were performed using the R programming language (version 2.7.2). The MIC score was computed using the Java binary provided by the authors, and its *P* values were determined using the precomputed tables available at the MINE web site. We conducted the multiple testing correction using the Benjamini–Hochberg procedure (Benjamini and Hochberg 1995) at a 5% of false discovery rate (FDR), which was implemented in the *multtest* package of the R programming language. We also used in-house developed Perl scripts for handling all genomic and expression data files.

Results

Gene Cluster Identification

We inferred a total of 31 conserved clusters that include both OBP and other nonhomologous genes (see Materials and Methods; table 1). These 31 clusters are maintained, on average, in 5.9 *Drosophila* species, comprise a mean of 8.3 genes and, more importantly, recover most of the OBP clusters defined in Vieira et al. (2007). For example, the cluster with highest gene density comprises four OBP genes (*Obp19a*, *Obp19b*, *Obp19c,* and *Obp19d*; cluster 1 in Vieira et al. [2007]) and one non-OBP gene in 7,330 bp. This cluster has been detected in 11 species, having a pBLS (cluster constraint probability) value of 0.995, and an adjusted pBLS (after correcting for the FDR [Benjamini and Hochberg 1995]) of 0.977 (table 1). In total, 14 of these clusters are significant (pBLS > 0.95), although only 10 remain after correcting for multiple testing (adjusted pBLS > 0.95). Therefore, these clusters are likely to be under functional constraints.

To determine specific features of the OBP gene organization, we compared clusters including OBP genes with all clusters identified in the *Drosophila* genomes. We inferred a total of 3,434 clusters (supplementary table S1, Supplementary Material online) that, on average, are conserved in 5.9 *Drosophila* species and encompass 6.4 genes (fig. 1). A total of 1,290 of the 3,434 clusters have a pBLS higher than 0.95, although only 58 remain significant after controlling for FDR. Because the FDR correction constitutes a conservative criterion (i.e., FDR methodologies reduce its statistical power as the number of tests increases [Carvajal-Rodriguez et al. 2009]), the actual number of clusters under functional constraint is likely to be higher than these 58 cases. Given that the raw pBLS value, which is not adjusted for multiple testing, is a continuous estimate of the cluster constraint strength, classifying clusters into significant and nonsignificant unbalanced categories will yield a further loss of statistical power (Pearson 1913). To avoid the negative effects of categorization, we analyzed the effect of competing factors on raw pBLS estimates using different association measures (supplementary table S2, Supplementary Material online), although only the values of the Spearman's rank correlation coefficient are reported throughout the manuscript.

Genes Clustered with OBP Genes Encode Plasma Membrane Proteins

We studied the existence of functional relationships among the genes clustered with OBPs by GO enrichment analysis (in total, 198 non-OBP genes). We compared the functionally annotated non-OBP genes in the 31 focal clusters (162 out of the 198 genes have GO annotations) with those present in all of the 3,434 *Drosophila* clusters (9,353 out of 11,811 genes). We found that the most characteristic GO terms among the genes clustered with OBPs are regulation of neurotransmitter transport, sodium channel activity, axon, neurotransmitter receptor activity, and integral to plasma membrane. After multiple testing correction (Benjamini and Hochberg 1995), only the latter category remained significant (hypergeometric test, $P = 1.34e{-}15$; table 2). As this analysis does not take into account the pBLS value of the clusters, we also separately reanalyzed the data from three different pBLS bins, each containing a similar number of genes. Notably, we found that the integral to plasma membrane GO term is enriched among the genes most conserving their neighborhood with the OBP genes.

The Cluster Conservation Correlates with the Type of Cis-Regulatory Elements

We analyzed the relevance of cis-regulatory elements in maintaining clusters including OBP genes. In particular, we

Table 1
The *Drosophila melanogaster* Clusters Including OBP Genes

	D. melanogaster Gene Cluster Region	No. of Genes	No. of OBPs	No. of Genomes Conserved	pBLS	Adjusted pBLS
Obp8a	X:9100153...9111401	4	1	8	0.925719	0.872071
Obp18a	X:19029114...19064675	3	1	2	0.733075	0.714666
Obp19a-d	X:20284679...20292009	5	4	11	0.995459	0.976539*
Obp22a	2L:1991705...2008966	4	1	4	0.734812	0.714666
Obp28a	2L:7426866...7497360	10	1	3	0.930950	0.874085
Obp44a	2R:4018938...4022588	2	1	12	0.921412	0.871778
Obp46a	2R:6194535...6209405	4	1	9	0.945767	0.887918
Obp47a	2R:6785747...6829206	4	1	5	0.893760	0.843170
Obp47b	2R:7189426...7197334	4	1	12	0.992088	0.964959*
Obp49a	2R:8574114...8645028	10	1	7	0.997471	0.983415*
Obp50a-c	2R:10257836...10260511	3	3	6	0.799992	0.753622
Obp50d	2R:10257836...10261264	4	1	5	0.793360	0.753622
Obp50e	2R:10262077...10299077	5	1	5	0.834610	0.786371
Obp51a	2R:10911880...10943746	2	1	4	0.603538	0.603538
Obp56a-c	2R:15585228...15588573	3	3	11	0.937764	0.879417
Obp56d-f	2R:15573111...15602373	9	3	3	0.895767	0.843170
Obp56g	2R:15656966...15671525	2	1	9	0.747767	0.714666
Obp56h	2R:15703059...15720473	2	1	10	0.840740	0.786371
Obp56i	2R:15703059...15768425	4	1	3	0.687717	0.676687
Obp57a-c	2R:16391061...16426819	10	3	4	0.951438	0.892469
Obp57d-e	2R:16413832...16449834	15	2	2	0.959350	0.903065
Obp58b-d; *Obp59a*	2R:18554661...18595219	11	4	5	0.988070	0.958908*
Obp69a	3L:12332216...12410803	7	1	9	0.990356	0.962628*
Obp73a	2R:5950890...6004962	6	1	9	0.986228	0.957306*
Obp76a	3L:19561538...19683092	20	1	3	0.999983	0.999483*
Obp83a-b	3R:1786045...1852962	6	2	4	0.839688	0.786371
Obp83cd; *Obp83ef*; *Obp83g*	3R:1880432...2129375	29	3	3	0.999967	0.999483*
Obp84a	3R:3050136...3113354	12	1	6	0.998575	0.985275*
Obp93a	3R:16774436...16966087	33	1	2	0.997325	0.983415*
Obp99a	3R:25456026...25501141	7	4	5	0.976460	0.933660
Obp99b-d	3R:25444756...25548111	17	3	2	0.97025	0.923146
Average		8.3	1.7	5.9		

NOTE.—The "no. of genes" and "no. of OBPs" columns indicate the total number of protein coding and OBP genes in the clusters, respectively. The "no. of genomes conserved" column represents the number of *Drosophila* species where the gene cluster region is identified.
*Significant clusters (adjusted pBLS > 0.95).

examined whether the pBLS value of such clusters is associated with the promoter architecture of the confined genes (i.e., the peaked or broad promoters as a proxy for the type of CPEs [Hoskins et al. 2011]). We found a significant correlation (Spearman's rank correlation coefficient: $\rho = 0.415$, $P = 0.044$; table 3), that is, the higher the pBLS value, the higher the proportion of broad-type promoters. Remarkably, this trend is also observed for all of the 3,434 *Drosophila* clusters (Spearman's rank correlation coefficient: $\rho = 0.044$, $P = 0.016$), indicating that gene clusters may have distinctive cis-regulatory elements.

The presence of the cis-regulatory elements shared among genes can restrict the movement of the target genes. For example, genes transcribed from shared promoters are common in many species, resulting in an excess of *head-to-head* gene pair arrangements (Trinklein et al. 2004; Kensche et al. 2008; Xu et al. 2009). We analyzed whether clusters including OBP genes have distinctive *head-to-head*, *tail-to-tail*, or *head-to-tail* gene pair organizations, but we detected no significant correlation with their pBLS value (Spearman's rank correlation coefficient, $P > 0.05$; supplementary fig. S1*A–C*, Supplementary Material online). In contrast, the results of the genome-wide analysis (including all 3,434 *Drosophila* clusters) were all significant (Spearman's rank correlation coefficient: $\rho = -0.095$, $P = 4.85e{-8}$; $\rho = 0.214$, $P < 2e{-16}$; $\rho = 0.110$, $P < 2.92e{-10}$ for the *head-to-tail*, *tail-to-tail* and *head-to-head* gene pair arrangements, respectively). Therefore, the sharing of cis-regulatory elements between contiguous genes is not a major factor in explaining the maintenance of OBP gene organization.

EB and EN Are Associated with the Conservation of Clusters That Include OBP Genes

As genes with broad-type promoters are often broadly expressed (Hoskins et al. 2011), we examined expression pattern

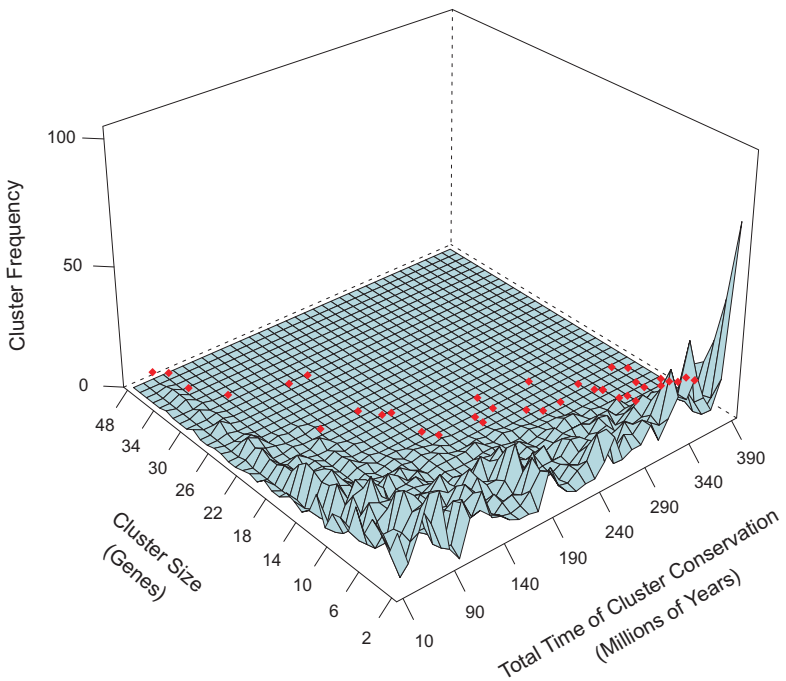


Fig. 1.—Frequency distribution of the 3,434 *Drosophila* clusters. Frequency distribution of the 3,434 *Drosophila* clusters, which is conditioned on the cluster size (i.e., number of genes per cluster) and the BLS value (total time of cluster conservation in million years ago). The 58 significant clusters after correcting for multiple testing are depicted in red.

Table 2
The 15 GO Terms Most Overrepresented among Genes Clustered with OBPs

GO Term	No. of Population Count	No. of Sample Count	*P* Value	Adjusted *P* Value
Integral to plasma membrane	180	22	6.78e−14	1.06e−10*
Sodium channel activity	35	4	0.0022	0.4634
GTPase activator activity	62	5	0.0030	0.4634
Retinal binding	6	2	0.0036	0.4634
Phototransduction	41	4	0.0040	0.4634
Metal ion transport	130	7	0.0047	0.4634
Monovalent inorganic cation transport	137	7	0.0062	0.4634
Locomotion	253	10	0.0071	0.4634
Neurotransmitter receptor activity	49	4	0.0075	0.4634
Locomotory behavior	144	7	0.0081	0.4634
Axon	52	4	0.0093	0.4634
Regulation of neurotransmitter secretion	10	2	0.0104	0.4634
Regulation of neurotransmitter transport	10	2	0.0104	0.4634
Sodium ion transport	56	4	0.0120	0.4634
Calcium-dependent phospholipid binding	11	2	0.0126	0.4634

Note.—The "Population Count" and "Sample Count" columns indicate the number of genes with GO annotation in the population (9,353 genes in the 3,434 *Drosophila* clusters) and sample (162 in genes clustered with OBPs), respectively. The "*P* value" column indicates the probability of observing such number of genes in the sample, given the number of genes in the population. *Overrepresented GO terms (adjusted $P < 0.05$).

Table 3
Summary of the Associations between pBLS and EB, EI, and EN

	OBP Clusters		Clusters with OBPs		All Clusters
	BC	PC	BC	PA	PA
EB	$\rho = 0.099$ ($P = 0.596$)	$t = 1.770$ ($P = 0.089$)	$\rho = 0.548$ ($P = 0.001$)	$\beta = 0.423$ ($P = 0.004$)	$\beta = 0.114$ ($P = 2.3e{-9}$)
EI	$\rho = -0.197$ ($P = 0.288$)	$t = -2.831$ ($P = 0.009$)	$\rho = 0.087$ ($P = 0.641$)	$\beta = -0.032$ ($P = 0.821$)	$\beta = 0.201$ ($P < 2e{-16}$)
EN	$\rho = 0.138$ ($P = 0.458$)	$t = 2.382$ ($P = 0.025$)	$\rho = 0.403$ ($P = 0.024$)	$\beta = 0.290$ ($P = 0.043$)	$\beta = 0.011$ ($P = 0.489$)

Note.—Relationship between pBLS and the EB, EI, and EN. The "OBP clusters," "Clusters with OBPs" and "All clusters" columns show results for clusters of OBP genes, for clusters including OBP genes, and for all 3,434 *Drosophila* clusters, respectively. "BC," "PC," and "PA" stand for bivariate correlation, partial correlation, and path analysis, respectively.

effects on cluster conservation. We found that the pBLS value of the clusters including OBP genes significantly correlates with EB (Spearman's rank correlation coefficient: $\rho = 0.548$, $P = 0.001$) and EN (Spearman's rank correlation coefficient: $\rho = 0.403$, $P = 0.024$), but not with EI (Spearman's rank correlation coefficient: $\rho = 0.087$, $P = 0.641$) (table 3). Nevertheless, these variables are highly intercorrelated: broadly expressed genes often exhibit high EI (Newman et al. 2006) and low EN (Lehner 2008). In addition, other factors, such as gene age (GA), may also hinder the causes of cluster conservation. For example, newly arising genes exhibit low EI and high gene loss rates (Wolf et al. 2009).

We determined the causal relationships among the factors involved in the OBP gene organization using path analysis (fig. 2), and assigning GA as the exogenous variable (i.e., not affected by factors of the underlying model). After factoring out the intercorrelated variables, EB ($\beta = 0.423$, $P = 0.004$) and EN ($\beta = 0.290$, $P = 0.043$) remained significant, that is, clusters including OBP genes are expressed in many tissues, exhibiting high stochastic fluctuations in transcript abundance, regardless of their EI ($\beta = -0.032$, $P = 0.821$). Interestingly, this result differs from the genome-wide analyses (3,434 clusters), where the pBLS value is affected by the EI ($\beta = 0.201$, $P < 2e{-16}$) and EB ($\beta = 0.114$, $P = 2e{-9}$), but not by the EN ($\beta = 0.011$, $P = 0.489$). However, the transcriptional effects on both data sets (including or not OBP genes) are not directly comparable, because they contain a different number and type of clusters. To evaluate whether EI and EN are specific features of the OBP gene organization, we thus performed computer simulations. We found that the EN effect (path coefficient from EN to pBLS) is higher for clusters including OBP genes than for random samples of 31 comparable clusters ($P = 0.035$), whereas the EI effect is lower ($P = 0.034$). Unlike comparable genome-wide clusters, clusters with OBP genes are not only influenced by EB but also by the EN, which does not support the clustering model of EN minimization.

OBP Genes in Conserved Clusters Also Exhibit Elevated Levels of EN

We analyzed whether the positive relationship between EN and cluster conservation remains significant after excluding non-OBP genes from the 31 conserved clusters. For that, we controlled for intercorrelated expression features. For example, we found that OBP genes in clusters with low pBLS, such as the *Obp22a* and *Obp50a* genes, are often transcribed in sexual tissues, which may suggest that the OBP gene organization has an SSE component (Spearman's rank correlation coefficient: $\rho = -0.420$, $P = 0.017$; fig. 3*A*). However, we found that this association is just a by-product of the OBP GA (partial correlation analysis, $t = -1.262$, $P = 0.219$; fig. 3*B*), supporting the observation that newly arising genes often exhibit an SSE pattern (Yeh et al. 2012). Actually, only the EI and EN of the OBP genes are directly associated with cluster conservation (partial correlation analysis, $t = -2.831$ and $t = 2.382$, $P = 0.009$ and $P = 0.025$, respectively). Overall, it supports the idea that EN may play a major role in shaping the OBP gene organization.

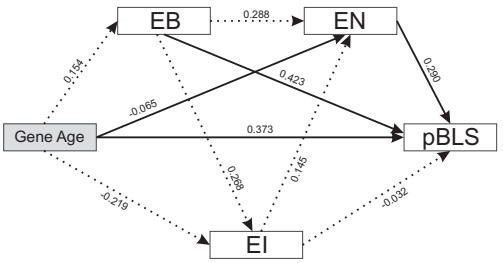


Fig. 2.—Transcriptional environment in clusters that include OBP genes. Path analysis model for the causal relationships among cluster constraint probability (pBLS), the minimum age of a gene in the cluster (GA), the EB, the EI, and the EN. The GA is the exogenous variable. The numbers on the lines indicate the path coefficients. Solid and dashed arrows represent significant and nonsignificant relationships.

Clusters Including OBP Genes Exhibit Distinctive Transcriptional Regulation by High-Order Chromatin Structures

We studied the effect of high-order chromatin structures (i.e., the nine specific chromatin states defined in Kharchenko et al. [2011]) on the conservation of clusters including OBP genes. We found a significant positive relationship between the pBLS

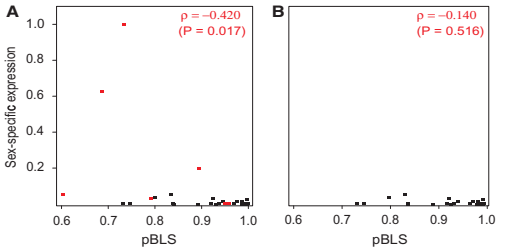


FIG. 3.—Genomic features of OBP genes. Relationship between pBLS and the SSE value using (*A*) all OBP genes and (*B*) after removing the recent OBP duplicates (red points).

value of these clusters and the proportion of nucleotides in the TE chromatin state (Spearman's rank correlation coefficient: $\rho = 0.480$, $P = 0.006$; fig. 4*A*). This chromatin state exhibits a distinct composition of proteins and histone marks (Kharchenko et al. 2011). As JIL-1 kinase is preferentially localized at the coding (Regnard et al. 2011) and promoter (Kellner et al. 2012) regions of the regulated genes, we analyzed its binding intensity separately for the coding, untranslated region, intergenic and intronic regions of the 31 focal clusters. We observed a strong positive correlation between the pBLS value and the JIL-1 binding intensity, though, after correcting by multiple testing, only remains statistically significant for the coding regions (Spearman's rank correlation coefficient: $\rho = 0.617$; $P = 2e{-}4$; fig. 4*B*). Taken together, these results suggest that the transcriptional regulation by high-order chromatin structures maintains the OBP gene organization to chromatin domains with the appropriate transcriptional environment (supplementary fig. S2, Supplementary Material online).

We further examined whether the high JIL-1 binding intensity and TE chromatin state represent particular features of clusters including OBP genes. Remarkably, the genome-wide cluster data set also shows significant correlation between the pBLS values and the JIL-1 binding intensities (Spearman's rank correlation coefficient: $\rho = 0.305$; $P < 2e{-}16$) and TE chromatin state (Spearman's rank correlation coefficient: $\rho = 0.312$; $P < 2e{-}16$). However, our computer simulations show that the correlation strengths are much higher for clusters including OBP than for random groups of 31 comparable clusters ($P < 1e{-}5$ and $P = 0.010$; fig. 4*C* and *D* for the TE chromatin state and for JIL-1), which suggests that the JIL-1 binding intensity and TE chromatin state are relevant factors explaining the conservation of clusters including OBP genes.

Discussion

Cluster Inference

Several methods have been developed to detect gene clusters conserved across a phylogeny (Lathe et al. 2000; Tamames 2001; Zheng et al. 2005). These methods differ in their underlying biological assumptions; therefore, their appropriateness depends on the biological question to be addressed. For example, the Synteny Database (Catchen et al. 2009) uses synteny information (i.e., it requires the same gene order and orientation across two genomes) to infer ortholog and paralog relationships, whereas the original version of the OperonDB algorithm (Ermolaeva et al. 2001) searches for clusters of physically close gene pairs conserved across different species to predict operons. The latter version of OperonDB (Pertea et al. 2009) improves the sensibility of the method by allowing rearrangement events inside the candidate cluster regions. There is compelling evidence indicating that some functional clusters can undergo internal rearrangements without transcriptional consequences (Itoh et al. 1999; Lathe et al. 2000); this observation led to the formation of the gene team model (Luc et al. 2003), which we applied here to infer *Drosophila* clusters.

Nevertheless, the gene team model implemented in the MCMuSeC software (Ling et al. 2009) also has some statistical problems. First, the inferred clusters can contain overlapping information, that is, a particular gene may be present in more than one cluster. Because such a feature violates the independence premise assumed for most statistical tests, we have confirmed that all of our conclusions hold after excluding overlapping clusters (1,634 out of 3,434 clusters have non-overlapping information, and 25 of these encompass OBP genes). Second, the statistical power to estimate conserved gene clusters increases with the species divergence time. Indeed, the 12 *Drosophila* species (Drosophila 12 Genomes 2007) used in this study are not divergent enough to detect small clusters (i.e., up to three genes). To detect such small clusters, it would be more appropriate to use more divergent species. However, the 12 *Drosophila* genomes provide a reasonable tradeoff between the quality of the assemblies and annotations (e.g., identification of orthologous and low sequence fragmentation in scaffolds) and the statistical power. This issue has important implications because two main classes of clusters have been described (Weber and Hurst 2011): small clusters of highly coexpressed genes (likely constrained by shared CREs) and large clusters of housekeeping and unrelated (i.e., nonhomologous) genes. Despite using genome data from 12 *Drosophila* species small-size clusters may be underestimated, this bias should not be relevant for the second cluster class. Thus, our results do not discard a relevant role of shared CREs in shaping genome architecture, but rather highlight the importance of high-order chromatin coregulatory mechanisms in the OBP gene organization. We mostly found large clusters (an average of 6.4 and 8.3 genes for clusters with and without OBP genes, respectively), which comprise a number of nonhomologous genes that exhibit high gene EB; features that characterize housekeeping gene clusters (i.e., the large-size cluster class).

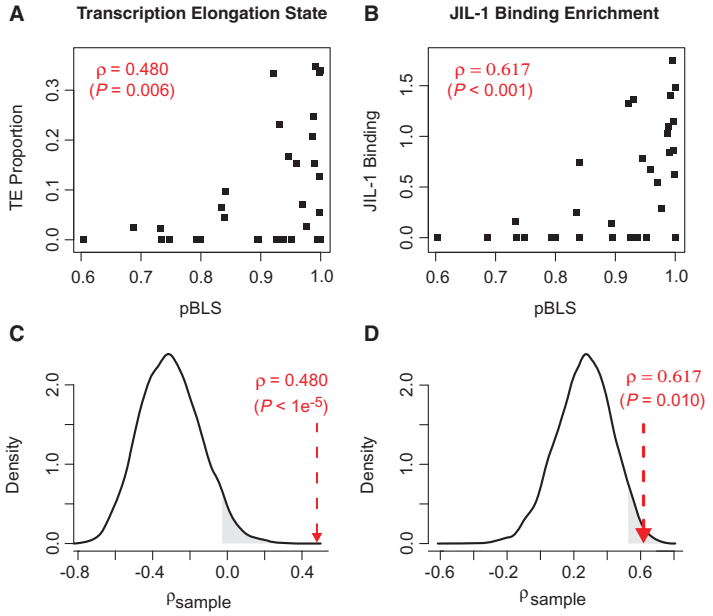


Fig. 4.—Chromatin features of the clusters that include OBP genes. Relationships between the cluster constraint probability (pBLS) and (*A*) the proportion of nucleotides annotated as TE and (*B*) JIL-1 binding intensity in coding regions. The ρ values are the correlation coefficients of these associations. Distribution of the correlation coefficients between pBLS values and (*C*) the proportion of TE and (*D*) JIL-1 binding intensities in *Drosophila* clusters obtained by computer simulations (10,000 replicates of 31 clusters). The arrow indicates the correlation coefficients observed for clusters including OBP genes ($P < 1e{-}5$ and $P = 0.010$, for the TE proportion and JIL-1 binding intensity, respectively). The shaded area in the right tail represents the 5% of the total distribution area.

Clusters Including OBP Genes Are Conserved by Functional Constraints

We identified 31 clusters including—at least—one OBP gene, and ten remained significant after correcting for multiple testing (table 1). Although natural selection may appear as the most immediate explanation for the conservation of the OBP genome organization, it could also represent a by-product of the uneven distribution of rearrangements along chromosomes (Ranz et al. 2001; Pevzner and Tesler 2003; Ruiz-Herrera et al. 2006; Bhutkar et al. 2008). Indeed, orthologous chromosome regions affected by a reduced number of rearrangements may maintain their cluster-like structure in the absence of functional constraints (von Grotthuss et al. 2010). However, such an explanation is unlikely to be the main reason for the maintenance of clusters including OBP genes. Indeed, homologous chromosome regions depleted in rearrangement breakpoints (and hence in rearrangements) are not common across *Drosophila* species (Ranz et al. 2001; Bhutkar et al. 2008; Schaeffer et al. 2008). In fact, the recombination rate, which is highly associated with the rearrangement rate, widely varies among closely related species (True et al. 1996). Consistently, we found no association between the recombination rate (Comeron et al. 2012) and pBLS values across the 31 focal clusters (Spearman's rank correlation coefficient: $\rho = -0.14$, $P = 0.47$), or across all 3,434 *Drosophila* clusters (Spearman's rank correlation coefficient: $\rho = 0.02$, $P = 0.16$) (supplementary fig. S3, Supplementary Material online). This lack of association results from the fact that we evaluated the statistical significance of the clusters using the observed divergence time of microsynteny conservation as null distribution. As this empirical null distribution depends upon the mode of chromosome evolution, it already captures the information of the uneven rearrangement distribution observed along *Drosophila* chromosomes. Therefore, it is unlikely that the OBP gene organization was a by-product of the rearrangement rate heterogeneity. In contrast, it may be constrained by natural selection for some functional meaning.

As conserved clusters of functionally or transcriptionally linked genes may include nonparalogous members, we defined a cluster as a group of genes that maintain their neighborhood across species regardless of whether they are homologous. This approximation is different from that used by Vieira and Rozas (2011) who only consider clusters of OBP paralogs. These authors observed that OBP genes are found

physically closer than expected by chance, although OR (odorant receptors) are not. In contrast, we found that some ORs are clustered with other nonhomologous genes (supplementary tables S1 and S3, Supplementary Material online). Similarly, clusters of OBP paralogs are conserved, but embedded within large arrangements that also include other non-OBP genes (table 1). For example, one of the most conserved *Drosophila* clusters includes *lush* (table 1 and fig. 5), which encodes an OBP involved in social aggregation and mating behavior (Xu et al. 2005), but also *Shal* (a potassium channel), *ash1* (involved in the ovoposition and oogenesis), *asf1* (dendrite morphogenesis), and *tey* (synaptic target recognition). Noticeably, the genes within this cluster also exhibit similar patterns of transcription across different developmental stages (fig. 5; Graveley et al. 2011). Overall, it suggests that some functional and transcriptional links maintain the *lush* genome cluster.

High-Order Chromatin Regulatory Mechanisms Provide the Appropriate Transcriptional Environment for Cluster Maintenance

Chromatin domains may restrict the location of genes to regions having the appropriate transcriptional environment (Noordermeer et al. 2011; Thomas et al. 2011), which may maintain the OBP gene organization. Nonhistone chromatin proteins regulating the chromatin state are therefore of particular interest. For example, lamin Dm0, which physically interacts with JIL-1 kinase (Bao et al. 2005), binds to gene clusters conserved across *Drosophila* species (Ranz et al. 2011). Remarkably, we found a strong association between the JIL-1 binding intensity and the maintenance of clusters including OBP genes (Spearman's rank correlation coefficient: $\rho = 0.617$, $P = 0.010$; fig. 4*D*). Moreover, genes regulated by JIL-1 kinase exhibit elevated levels of EB (Regnard et al. 2011) and EN (JIL-1 releases the paused RNA polymerase II at the proximal-promoter (Kellner et al. 2012), favoring transcriptional elongation bursts that increase EN [Becskei et al. 2005; Kaern et al. 2005; Rajala et al. 2010]). Consistent with this idea, we have shown that the OBP gene organization is associated with elevated levels of EB ($P = 0.004$) and EN ($P = 0.043$).

It has been shown that housekeeping genes may be particularly confined to chromosome regions possessing the appropriate transcriptional environment; indeed, mutations that alter their location may exert important deleterious pleiotropic effects in diverse tissues and developmental stages (Wang and Zhang 2010). Batada and Hurst (2007) have suggested that broadly expressed genes are located in chromosome regions with low stochastic transcriptional fluctuations to minimize the deleterious effects of EN. However, the functional constraints underlying the conservation of the OBP gene organization do not support this hypothesis. First, clusters with OBP genes often exhibit a high proportion of broad-type promoters, which yield elevated levels of EB. Although these two features (broad-promoters and EB) are associated with reduced levels of EN (Tirosh and Barkai 2008; Wang and Zhang 2010; Xi et al. 2011), we detected a positive relationship between the stochastic transcriptional fluctuation (EN)

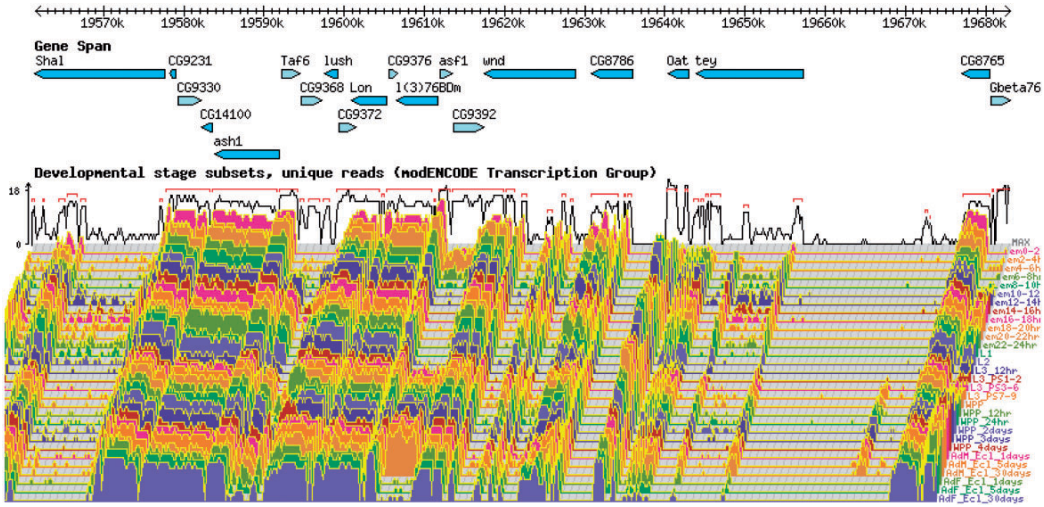


Fig. 5.—The cluster including the *lush* (*Obp76a*) gene. The cluster (pBLS value of 0.999983) including *lush* (*Obp76a*) and other 19 non-OBP genes (blue boxes). The coordinates (from 19,570 k to 19,680 k) correspond to the 3L chromosome of *Drosophila melanogaster*. The intensity peaks below the genes indicate the EI values across 30 developmental stages (in different colors).

and the pBLS value of these clusters (table 3). Second, although EN can be alleviated by increasing EI (Lehner 2008), the most conserved clusters include OBP genes not only with the highest EN (partial correlation analysis, $P = 0.025$) but also with the lowest EI (partial correlation analysis; $P = 0.009$; table 3). In fact, the EI effect on pBLS is lower for clusters with OBP genes than for random samples of 31 comparable clusters ($P = 0.034$). Finally, even though *head-to-head* gene pair arrangements can minimize EN (Wang et al. 2011), clusters with OBP genes do not exhibit a significant correlation between the pBLS value and the proportion of *head-to-head* gene pair frequency. Therefore, a suitable transcriptional environment need not always have reduced levels of EN; indeed, a clustering model based on elevated EN levels may explain the OBP gene organization.

Some theoretical models predict that, under certain circumstances, EN can even be beneficial as a source for natural variation, particularly for proteins acting in changing environments (e.g., stress response proteins such as oxidative kinases [Dong et al. 2011]). Some empirical results are consistent with this model. In yeast, for example, the elevated EN of plasma-membrane transporters appears to be driven by positive selection (Zhang et al. 2009). The genes clustered with OBPs also encode membrane proteins and, interestingly, many of these proteins have transporter activity (table 2). In fact, the extensive transcriptional diversification of the OBPs suggests that, apart from transporting odorants of the external environment, some OBPs also act as general carriers of hydrophobic molecules through the extracellular matrix (Arya et al. 2010). Therefore, higher EN levels may allow for the detection of wider ranges of concentrations of hydrophobic molecules. Fluctuations in OBP transcript abundance may represent an important mechanism to increase phenotypic plasticity. Mutations affecting OBP mRNA stability (Wang et al. 2007) and reduced OBP expression levels (Swarup et al. 2011) can actually elicit different *Drosophila* behaviors to particular odorants, that is, fluctuations in OBP transcript abundance can play a key role in the combinatorial nature of the olfactory coding process. Therefore, natural selection may have favored assembling OBP genes in chromosomal regions with high EN, which in turn may have led to the observed structure of OBP genes in clusters of functionally and transcriptionally related genes.

Supplementary Material

Supplementary tables S1–S3 and figures S1–S3 are available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

Acknowledgments

J.R. conceived and supervised all research. P.L. developed the bioinformatics tools, analyzed the data, and wrote the first version of the manuscript. Both the authors approved the final manuscript. The authors thank J.M. Ranz, F.G. Vieira, and three anonymous reviewers for their comments and suggestions on the manuscript. This work was supported the Ministerio de Ciencia e Innovación of Spain grants BFU2007-62927 and BFU2010-15484, the Comissió Interdepartamental de Recerca I Innovació Tecnològica of Spain grant 2009SGR-1287, and the ICREA Acadèmia (Generalitat de Catalunya) grant to J.R. (partially supported).

Literature Cited

Arya GH, et al. 2010. Natural variation, functional pleiotropy and transcriptional contexts of odorant binding protein genes in *Drosophila melanogaster*. Genetics 186:147–1485.

Bao X, et al. 2005. The JIL-1 kinase interacts with lamin Dm0 and regulates nuclear lamina morphology of *Drosophila* nurse cells. J Cell Sci. 118: 5079–5087.

Batada NN, Hurst LD. 2007. Evolution of chromosome organization driven by selection for reduced gene expression noise. Nat Genet. 39: 945–949.

Bauer S, Grossmann S, Vingron M, Robinson PN. 2008. Ontologizer 2.0—a multifunctional tool for GO term enrichment analysis and data exploration. Bioinformatics 24:1650–1651.

Becskei A, Kaufmann BB, van Oudenaarden A. 2005. Contributions of low molecule number and chromosomal positioning to stochastic gene expression. Nat Genet. 37:937–944.

Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J Royal Stat Soc B. 57:289–300.

Bhutkar A, et al. 2008. Chromosomal rearrangement inferred from comparisons of 12 *Drosophila* genomes. Genetics 179:1657–1680.

Biessmann H, et al. 2010. The *Anopheles gambiae* odorant binding protein 1 (AgamOBP1) mediates indole recognition in the antennae of female mosquitoes. PLoS One 5:e9471.

Boutanaev AM, Kalmykova AI, Shevelyov YY, Nurminsky DI. 2002. Large clusters of co-expressed genes in the *Drosophila* genome. Nature 420: 666–669.

Capelson M, et al. 2010. Chromatin-bound nuclear pore components regulate gene expression in higher eukaryotes. Cell 140:372–383.

Caron H, et al. 2001. The human transcriptome map: clustering of highly expressed genes in chromosomal domains. Science 291: 1289–1292.

Carvajal-Rodriguez A, de Una-Alvarez J, Rolan-Alvarez E. 2009. A new multitest correction (SGoF) that increases its statistical power when increasing the number of tests. BMC Bioinformatics 10:209.

Catchen JM, Conery JS, Postlethwait JH. 2009. Automated identification of conserved synteny after whole-genome duplication. Genome Res. 19:1497–1505.

Comeron JM, Ratnappan R, Bailin S. 2012. The many landscapes of recombination in *Drosophila melanogaster*. PLoS Genet. 8: e1002905.

Chamaon K, Smalla KH, Thomas U, Gundelfinger ED. 2002. Nicotinic acetylcholine receptors of *Drosophila*: three subunits encoded by genomically linked genes can co-assemble into the same receptor complex. J Neurochem. 80:149–157.

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

Dong D, Shao X, Deng N, Zhang Z. 2011. Gene expression variations are predictive for stochastic noise. Nucleic Acids Res. 39:403–413.

Drosophila 12 Genomes C, et al. 2007. Evolution of genes and genomes on the *Drosophila* phylogeny. Nature 450:203–218.

Engstrom PG, Ho Sui SJ, Drivenes O, Becker TS, Lenhard B. 2007. Genomic regulatory blocks underlie extensive microsynteny conservation in insects. Genome Res. 17:1898–1908.
Ermolaeva MD, White O, Salzberg SL. 2001. Prediction of operons in microbial genomes. Nucleic Acids Res. 29:1216–1221.
Filion GJ, et al. 2010. Systematic protein location mapping reveals five principal chromatin types in *Drosophila* cells. Cell 143:212–224.
Foret S, Maleszka R. 2006. Function and evolution of a gene family encoding odorant binding-like proteins in a social insect, the honey bee (*Apis mellifera*). Genome Res. 16:1404–1413.
Fraser HB, Hirsh AE, Giaever G, Kumm J, Eisen MB. 2004. Noise minimization in eukaryotic gene expression. PLoS Biol. 2:e137.
Gan M, Moebus S, Eggert H, Saumweber H. 2011. The Chriz-Z4 complex recruits JIL-1 to polytene chromosomes, a requirement for interband-specific phosphorylation of H3S10. J Biosci. 36:425–438.
Gilbert N, et al. 2004. Chromatin architecture of the human genome: gene-rich domains are enriched in open chromatin fibers. Cell 118: 555–566.
Graveley BR, et al. 2011. The developmental transcriptome of *Drosophila melanogaster*. Nature 471:473–479.
Grossmann S, Bauer S, Robinson PN, Vingron M. 2007. Improved detection of overrepresentation of gene-ontology annotations with parent child analysis. Bioinformatics 23:3024–3031.
Hekmat-Scafe DS, Scafe CR, McKinney AJ, Tanouye MA. 2002. Genome-wide analysis of the odorant-binding protein gene family in *Drosophila melanogaster*. Genome Res. 12:1357–1369.
Hoskins RA, et al. 2011. Genome-wide analysis of promoter architecture in *Drosophila melanogaster*. Genome Res. 21:182–192.
Huerta-Cepas J, Gabaldon T. 2011. Assigning duplication events to relative temporal scales in genome-wide studies. Bioinformatics 27:38–45.
Itoh T, Takemoto K, Mori H, Gojobori T. 1999. Evolutionary instability of operon structures disclosed by sequence comparisons of complete microbial genomes. Mol Biol Evol. 16:332–346.
Jin Y, et al. 1999. JIL-1: a novel chromosomal tandem kinase implicated in transcriptional regulation in *Drosophila*. Mol Cell. 4:129–135.
Kaern M, Elston TC, Blake WJ, Collins JJ. 2005. Stochasticity in gene expression: from theories to phenotypes. Nat Rev Genet. 6:451–464.
Kalmykova AI, Nurminsky DI, Ryzhov DV, Shevelyov YY. 2005. Regulated chromatin domain comprising cluster of co-expressed genes in *Drosophila melanogaster*. Nucleic Acids Res. 33:1435–1444.
Kellner WA, Ramos E, Van Bortle K, Takenaka N, Corces VG. 2012. Genome-wide phosphoacetylation of histone H3 at *Drosophila* enhancers and promoters. Genome Res. 22:1081–1088.
Kensche PR, Oti M, Dutilh BE, Huynen MA. 2008. Conservation of divergent transcription in fungi. Trends Genet. 24:207–211.
Kharchenko PV, et al. 2011. Comprehensive analysis of the chromatin landscape in *Drosophila melanogaster*. Nature 471:480–485.
Krieger MJ, Ross KG. 2002. Identification of a major gene regulating complex social behavior. Science 295:328–332.
Kruse SW, Zhao R, Smith DP, Jones DNM. 2003. Structure of a specific alcohol-binding site defined by the odorant binding protein LUSH from *Drosophila melanogaster*. Nat Struct Mol Biol. 10:694.
Lathe WC 3rd, Snel B, Bork P. 2000. Gene context conservation of a higher order than operons. Trends Biochem Sci. 25:474–479.
Laughlin JD, Ha TS, Jones DNM, Smith DP. 2008. Activation of pheromone-sensitive neurons is mediated by conformational activation of pheromone-binding protein. Cell 133:1255–1265.
Lee JM, Sonnhammer EL. 2003. Genomic gene clustering analysis of pathways in eukaryotes. Genome Res. 13:875–882.
Lehner B. 2008. Selection to minimise noise in living systems and its implications for the evolution of gene expression. Mol Syst Biol. 4:170.
Lercher MJ, Urrutia AO, Hurst LD. 2002. Clustering of housekeeping genes provides a unified model of gene order in the human genome. Nat Genet. 31:180–183.
Li G, Reinberg D. 2011. Chromatin higher-order structures and gene regulation. Curr Opin Genet Dev. 21:175–186.
Li YY, et al. 2006. Systematic analysis of head-to-head gene organization: evolutionary conservation and potential biological relevance. PLoS Comput Biol. 2:e74.
Librado P, Vieira FG, Rozas J. 2012. BadiRate: estimating family turnover rates by likelihood-based methods. Bioinformatics 28:279–281.
Ling X, He X, Xin D. 2009. Detecting gene clusters under evolutionary constraint in a large number of genomes. Bioinformatics 25:571–577.
Luc N, Risler JL, Bergeron A, Raffinot M. 2003. Gene teams: a new formalization of gene clusters for comparative genomics. Comput Biol Chem. 27:59–67.
Luz H, Staub E, Vingron M. 2006. About the interrelation of evolutionary rate and protein age. Genome Inform. 17:240–250.
Maeda RK, Karch F. 2007. Making connections: boundaries and insulators in *Drosophila*. Curr Opin Genet Dev. 17:394–399.
Matsuo T, Sugaya S, Yasukawa J, Aigaki T, Fuyama Y. 2007. Odorant-binding proteins OBP57d and OBP57e affect taste perception and host-plant preference in *Drosophila sechellia*. PLoS Biol. 5:e118.
Negre N, et al. 2010. A comprehensive map of insulator elements for the *Drosophila* genome. PLoS Genet. 6:e1000814.
Newman JR, et al. 2006. Single-cell proteomic analysis of *S. cerevisiae* reveals the architecture of biological noise. Nature 441:840–846.
Noordermeer D, et al. 2011. The dynamic architecture of Hox gene clusters. Science 334:222–225.
Pearson K. 1913. On the measurement of the influence of "broad categories" on correlation. Biometrika 9:116–139.
Pertea M, Ayanbule K, Smedinghoff M, Salzberg SL. 2009. OperonDB: a comprehensive database of predicted operons in microbial genomes. Nucleic Acids Res. 37:D479–D482.
Pevzner P, Tesler G. 2003. Human and mouse genomic sequences reveal extensive breakpoint reuse in mammalian evolution. Proc Natl Acad Sci U S A. 100:7672–7677.
Rach EA, et al. 2011. Transcription initiation patterns indicate divergent strategies for gene regulation at the chromatin level. PLoS Genet. 7: e1001274.
Rajala T, Hakkinen A, Healy S, Yli-Harja O, Ribeiro AS. 2010. Effects of transcriptional pausing on gene expression dynamics. PLoS Comput Biol. 6:e1000704.
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.
Ranz JM, Diaz-Castillo C, Petersen R. 2011. Conserved gene order at the nuclear periphery in *Drosophila*. Mol Biol Evol. 29:13–16.
Rath U, et al. 2006. The chromodomain protein, Chromator, interacts with JIL-1 kinase and regulates the structure of *Drosophila* polytene chromosomes. J Cell Sci. 119:2332–2341.
Regnard C, et al. 2011. Global analysis of the relationship between JIL-1 kinase and transcription. PLoS Genet. 7:e1001327.
Reshef DN, et al. 2011. Detecting novel associations in large data sets. Science 334:1518–1524.
Ruiz-Herrera A, Castresana J, Robinson TJ. 2006. Is mammalian chromosomal evolution driven by regions of genome fragility? Genome Biol. 7:R115.
Sanchez-Gracia A, Rozas J. 2011. Molecular population genetics of the OBP83 genomic region in *Drosophila subobscura* and *D. guanche*: contrasting the effects of natural selection and gene arrangement expansion in the patterns of nucleotide variation. Heredity 106:191–201.
Schaeffer SW, et al. 2008. Polytene chromosomal maps of 11 *Drosophila* species: the order of genomic scaffolds inferred from genetic and physical maps. Genetics 179:1601–1655.
Swarup S, Williams TI, Anholt RR. 2011. Functional dissection of odorant binding protein genes in *Drosophila melanogaster*. Genes Brain Behav. 10:648–657.

Tamames J. 2001. Evolution of gene order conservation in prokaryotes. Genome Biol. 2:RESEARCH0020.
Tamura K, Subramanian S, Kumar S. 2004. Temporal patterns of fruit fly (*Drosophila*) evolution revealed by mutation clocks. Mol Biol Evol. 21: 36–44.
Tegoni M, Campanacci V, Cambillau C. 2004. Structural aspects of sexual attraction and chemical communication in insects. Trends Biochem Sci. 29:257–264.
Thomas S, et al. 2011. Dynamic reprogramming of chromatin accessibility during *Drosophila* embryo development. Genome Biol. 12:R43.
Tirosh I, Barkai N. 2008. Two strategies for gene regulation by promoter nucleosomes. Genome Res. 18:1084–1091.
Trinklein ND, et al. 2004. An abundance of bidirectional promoters in the human genome. Genome Res. 14:62–66.
True JR, Mercer JM, Laurie CC. 1996. Differences in crossover frequency and distribution among three sibling species of *Drosophila*. Genetics 142:507–523.
Vaquerizas JM, et al. 2010. Nuclear pore proteins nup153 and megator define transcriptionally active regions in the *Drosophila* genome. PLoS Genet. 6:e1000846.
Vieira FG, Rozas J. 2011. Comparative genomics of the odorant-binding and chemosensory protein gene families across the Arthropoda: origin and evolutionary history of the chemosensory system. Genome Biol Evol. 3:476–490.
Vieira FG, Sanchez-Gracia A, Rozas J. 2007. Comparative genomic analysis of the odorant-binding protein family in 12 *Drosophila* genomes: purifying selection and birth-and-death evolution. Genome Biol. 8: R235.
von Grotthuss M, Ashburner M, Ranz JM. 2010. Fragile regions and not functional constraints predominate in shaping gene organization in the genus *Drosophila*. Genome Res. 20:1084–1096.
Wallace HA, Plata MP, Kang HJ, Ross M, Labrador M. 2009. Chromatin insulators specifically associate with different levels of higher-order chromatin organization in *Drosophila*. Chromosoma 119:177–194.
Wang GZ, Lercher MJ, Hurst LD. 2010. Transcriptional coupling of neighboring genes and gene expression noise: evidence that gene orientation and noncoding transcripts are modulators of noise. Genome Biol Evol. 3:320–331.
Wang GZ, Lercher MJ, Hurst LD. 2011. Transcriptional coupling of neighboring genes and gene expression noise: evidence that gene orientation and noncoding transcripts are modulators of noise. Genome Biol Evol. 3:320–331.
Wang P, Lyman RF, Shabalina SA, Mackay TF, Anholt RR. 2007. Association of polymorphisms in odorant-binding protein genes with variation in olfactory response to benzaldehyde in *Drosophila*. Genetics 177:1655–1665.
Wang Z, Zhang J. 2010. Impact of gene expression noise on organismal fitness and the efficacy of natural selection. Proc Natl Acad Sci U S A. 108:E67–E76.
Weber CC, Hurst LD. 2011. Support for multiple classes of local expression clusters in *Drosophila melanogaster*, but no evidence for gene order conservation. Genome Biol. 12:R23.
Wolf YI, Novichkov PS, Karev GP, Koonin EV, Lipman DJ. 2009. The universal distribution of evolutionary rates of genes and distinct characteristics of eukaryotic genes of different apparent ages. Proc Natl Acad Sci U S A. 106:7273–7280.
Xi Y, Yao J, Chen R, Li W, He X. 2011. Nucleosome fragility reveals novel functional states of chromatin and poises genes for activation. Genome Res. 21:718–724.
Xu P, Atkinson R, Jones DN, Smith DP. 2005. *Drosophila* OBP LUSH is required for activity of pheromone-sensitive neurons. Neuron 45: 193–200.
Xu Z, et al. 2009. Bidirectional promoters generate pervasive transcription in yeast. Nature 457:1033–1037.
Yang L, Yu J. 2009. A comparative analysis of divergently-paired genes (DPGs) among *Drosophila* and vertebrate genomes. BMC Evol Biol. 9: 55.
Yeh S-D, et al. 2012. Functional evidence that a recently evolved *Drosophila* sperm-specific gene boosts sperm competition. Proc Natl Acad Sci U S A. 109:2043–2048.
Zhang Z, Qian W, Zhang J. 2009. Positive selection for elevated gene expression noise in yeast. Mol Syst Biol. 5:299.
Zheng Y, Anton BP, Roberts RJ, Kasif S. 2005. Phylogenetic detection of conserved gene clusters in microbial genomes. BMC Bioinformatics 6: 243.
Zhou S, Stone EA, Mackay TF, Anholt RR. 2009. Plasticity of the chemoreceptor repertoire in *Drosophila melanogaster*. PLoS Genet. 5: e1000681.

Associate editor: Gunter Wagner

Supplementary Material

Uncovering the Functional Constraints Underlying the Genomic Organization of the Odorant-Binding Protein Genes

P. Librado and J. Rozas

Departament de Genètica and Institut de Recerca de la Biodiversitat,

Universitat de Barcelona, Diagonal 643, 08028 Barcelona, Spain.

Table S3. The *D. melanogaster* OR clusters

OR	*D. melanogaster* Gene Cluster Region	#Genes	#ORs	#Genomes conserved	pBLS	Adjusted pBLS
Or1a	X:201717..364670	3	1	10	0.9312	0.8781
Or2a	X:2119744..2169111	12	1	3	0.9922	0.9602*
Or7a	X:7972382..8071187	26	1	3	0.9975	0.9731*
Or9a	X:10352219..10372526	5	1	6	0.9507	0.8991
Or10a	X:11307533..11348335	5	1	8	0.9699	0.9238
Or22a-b	2L:1492295..1524259	10	2	2	0.8999	0.8565
Or22c	2L:2036456..2178749	14	1	2	0.9371	0.8781
Or23a	2L:2646577..2815760	29	1	2	0.9967	0.9718*
Or24a	2L:4164689..4345776	12	1	2	0.9197	0.8721
Or30a	2L:9111814..9247187	5	1	5	0.9132	0.8667
Or33a-c	2L:11934505..11948634	5	3	5	0.8346	0.8077
Or35a	2L:15613948..15629552	5	1	6	0.9336	0.8781
Or42a-b	2R:1645560..1686017	6	2	4	0.8385	0.8077
Or43a	2R:3120344..3200855	8	1	6	0.9928	0.9602*
Or43b	2R:3802532..3818521	6	1	8	0.9701	0.9238
Or45a	2R:5205046..5305351	11	1	2	0.8884	0.8452
Or45b	2R:5447103..5451378	2	1	10	0.7979	0.7772
Or46a	2R:5938367..6079899	23	1	2	0.9889	0.9568*
Or47a	2R:7137873..7162796	9	1	4	0.9707	0.9238
Or47b	2R:7207215..7232620	8	1	4	0.9656	0.9179
Or49a	2R:8728318..8786900	12	1	7	0.9943	0.9602*
Or56a	2R:15656966..15671525	2	1	9	0.7478	0.7355
Or59a	2R:19273201..19328588	8	1	7	0.9915	0.9602*
Or59b-c	2R:19357925..19363941	3	2	7	0.8457	0.8077
Or63a	3L:2883638..3088493	21	1	2	0.9846	0.9526*
Or67a	3L:9507219..9588339	4	1	3	0.7162	0.7094
Or67b	3L:9590954..9674160	15	1	2	0.9550	0.9032
Or67c-d	3L:10158909..10267724	8	2	6	0.9849	0.9526*
Or69a	3L:12945783..13017862	14	1	4	0.9846	0.9526*
Or74a	3L:17315941..17332358	4	1	7	0.8844	0.8446
Or82a	3R:80502..84166	2	1	12	0.9214	0.8721
Or83a-b	3R:1219650..1290472	7	2	8	0.9985	0.9781*
Or83c	3R:1880432..2129375	29	1	3	0.9999	0.9986*
Or85a	3R:4103514..4189616	29	1	3	0.9988	0.9781*
Or85c	3R:4335799..4351846	5	1	2	0.7051	0.7051
Or85d	3R:4378330..4389170	4	1	8	0.8999	0.8565
Or85f	3R:5189032..5225273	4	1	8	0.9291	0.8781
Or88a	3R:9882690..9977721	15	1	2	0.9593	0.9080
Or92a	3R:16204718..16338169	7	1	2	0.7691	0.7518
Or98a	3R:23671672..23745195	6	1	2	0.8357	0.8077
Or98b	3R:24117060..24179224	8	1	2	0.8301	0.8077

Or99a-b	3R:18802993..18819283	4	2	3	0.8318	0.8077
Average		9.9	1.2	4.8		

The '#Genes'and '#ORs' columns indicate the total number of coding and OR genes in the cluster regions, respectively. The '#Genomes conserved' column the number of *Drosophila* species where the gene cluster region is conserved. Asterisks show significant OR clusters (adjusted pBLS > 0.95).

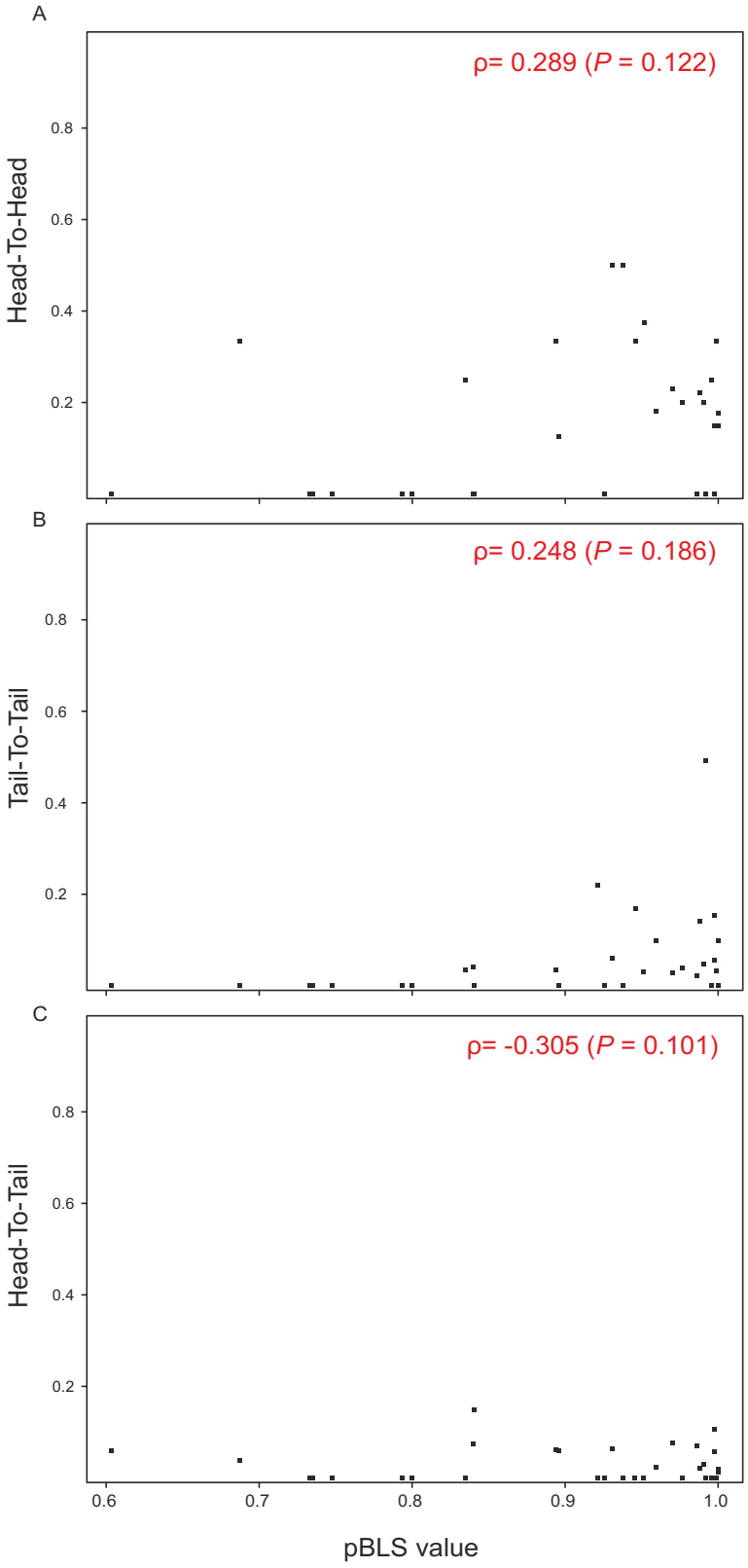


Figure S1. Correlation between the cluster constraint probability (pBLS) and the proportion of (A) Head-To-Head, (B) Tail-To-Tail and (C) Head-To-Tail gene pair arrangements in the OBP clusters.

3.5 Positive selection drives the evolution of the transcriptional regulatory upstream regions of the major chemosensory gene families

El sistema quimiosensorial de los animales está involucrado en procesos biológicos críticos para su supervivencia, muchos de los cuales están mediados por familias multigénicas. Varios estudios han demostrado que las regiones codificadoras de los genes del sistema quimiosensorial han evolucionado, principalmente, por selección purificadora. No obstante, también se han detectado algunas posiciones con la huella molecular de la selección positiva. El conocimiento evolutivo de las regiones codificadoras, contrasta con la falta de estudios en sus regiones *upstream*, especialmente dadas las evidencias que vinculan ciertos cambios transcripcionales con efectos fenotípicos.

En esta tesis doctoral, hemos cuantificado la contribución relativa de la selección natural a la evolución molecular de las regiones *upstream* de los genes quimiosensoriales. Para el análisis, hemos integrado datos del proyecto "Drosophila Genetic Reference Panel" (DGRP) y de la anotación funcional de los elementos *cis*-reguladores (CREs). Los resultados muestran que la selección natural tiene un impacto significativo en la evolución de las secuencias *upstream*, tanto a nivel de divergencia nucleotídica, como a nivel de ganancia y muerte de CREs, un resultado que es especialmente pronunciado en las familias que codifican para receptores olfativos (ORs) y gustativos (GRs).

Positive selection drives the evolution of the transcriptional regulatory upstream regions of the major chemosensory gene families

Pablo Librado and Julio Rozas[1]

Departament de Genètica and Institut de Recerca de la Biodiversitat (IRBio), Universitat de Barcelona, Av. Diagonal 643, Barcelona 08028, Spain

[1]Corresponding author

Pablo Librado: plibrado@ub.edu; Julio Rozas: jrozas@ub.edu

Abstract

The animal chemosensory system is involved in critical and essential biological processes, most of them mediated by proteins encoded in multigene families. Several studies have revealed that the chemosensory protein-coding regions evolve under purifying selection; even so a few positions show the molecular hallmark of positive selection. This relatively well-defined mode of protein-coding evolution contrasts with the lack of studies at their upstream regions, especially taking into account the large body of evidence linking transcriptional changes at the chemosensory genes and phenotypic effects.

Here, we quantified the relative contribution of natural selection to the molecular evolution of the upstream regions of the major chemosensory gene families. For the analyses, we have integrated data from the Drosophila Genetic Reference Panel (DGRP) project, altogether with the functional annotation of the fly *cis*-regulatory elements (CREs). We show that natural selection has played a major role in driving the evolution of the chemosensory upstream regions at the level of both, sequence divergence and CRE gain and loss, especially for the OR and GR gene families.

Introduction

The chemosensory system is involved in essential processes such as nutrition, reproduction and social communication (Krieger and Ross 2002; Matsuo et al. 2007; Whiteman and Pierce 2008; Smadja and Butlin 2009). In insects, the specificity and sensibility of this system highly rely on the peripheral processes occurring in sensilla (Pelosi 1996; Hildebrand and Shepherd 1997): the uptake of chemical stimuli, its transport through the sensilla lymph, and the interaction with the chemoreceptor that will activate the signal transduction cascade. Most of these peripheral processes are mediated by proteins encoded in multigene families (Sanchez-Gracia et al. 2009), including extracellular ligand-binding proteins, such as Odorant Binding Proteins (OBPs) and Chemosensory Proteins (CSPs), and membrane receptor proteins, such as Odorant Receptors (ORs), Gustatory Receptors (GRs) and Ionotropic Receptors (IRs). Since the correct expression of such genes determines the ability to discriminate external chemical cues (and thus the individual fitness), the chemosensory system constitutes an excellent candidate to study the role of natural selection in driving transcriptional regulation at the molecular level.

Several studies have revealed that the chemosensory protein-coding regions evolve under purifying selection (for a review, Sanchez-Gracia et al. 2009). Even so, the selective constraint is lower for the 'odorant binding' functional category than for the genome average (Clark et al. 2007), and a few positions show the molecular hallmark of positive selection (Foret and Maleszka 2006; McBride et al. 2007; Sánchez-Gracia and Rozas 2008). This relatively well-defined mode of protein-coding gene evolution contrasts with the lack of studies surveying the upstream regions of the chemosensory genes (hereafter, chemosensory upstream regions). Indeed, there are compelling

evidence linking transcriptional changes and phenotypic effects. For example: (i) expression changes in the *Or10* (Rollmann et al. 2010)*, Or43a* (Stortkuhl et al. 2005), and *Obp99a* (Wang et al. 2010) have been associated with behavioural responses to benzaldehyde, an odorant involved in the *D. melanogaster* feeding behaviour; (ii) the transcription of several OBP and OR genes, such as the *Or22a* and the *Obp50a*, is upregulated in *D. sechellia*, likely reflecting its host specialization (Kopp et al. 2008); (iii) the OBP genes are often located in chromosome regions with a transcriptional background that may be advantageous in unstable external environments (Librado and Rozas 2013). Overall, these findings support the critical involvement of the chemosensory transcriptional changes in driving the phenotypic variation, both within (polymorphism) and between (divergence) species.

Comparison of polymorphism and divergence variability patterns between 5' upstream and neutrally evolving regions can be instrumental for understanding the action of natural selection on regulatory sequences (Jenkins et al. 1995; Egea et al. 2008). This powerful approach allows -in addition- estimating the fraction of substitutions fixed by positive selection (*i.e.* the so-called α parameter (Fay et al. 2001)). Noticeably, α estimates in non-coding regions widely vary among species, ranging from ~0% in humans (Keightley et al. 2005; Eyre-Walker and Keightley 2009) to ~50% in the UTRs of *Drosophila* (Andolfatto 2005; Haddrill et al. 2008). Nevertheless, this variation need not necessarily indicate differences in the adaptive pressures, but may only reflect the reduced effectiveness of natural selection in small populations (*i.e.* with low effective population size, or N_e) (Jensen and Bachtrog 2011; Gossmann et al. 2012). Recently, Schneider et *al.* (2011) developed a new approach that first accounts for N_e variations over time, and then estimates the main selective parameters: the distribution of fitness

effects of new deleterious mutations (DFE), the proportion of adaptive mutations (p_a) with a certain selective strength (s_a), the rate of adaptive substitution (α), and the ratio of adaptive-to-neutral substitution rates (ω_α).

Here, we examined the relative contribution of natural selection to the molecular evolution of the 5' upstream gene regions of the major chemosensory families. For that, we have integrated data from the Drosophila Genetic Reference Panel (DGRP) project (Mackay et al. 2012), altogether with the functional annotation of the fly genome (Roy et al. 2010). We show that natural selection has played a major role in driving the evolution of the 5' chemosensory upstream regions, especially in the sequence divergence and turnover of their *cis*-regulatory elements (CREs). We show that natural selection has played a major role in driving the evolution of the chemosensory upstream regions at the level of both, sequence divergence and CRE gain and loss, especially for the OR and GR gene families.

Material and Methods

Genomic alignments

We downloaded the genome-wide multiple sequence alignment (MSA) of 158 *D. melanogaster* (Mackay et al. 2012), with *D. simulans* and *D. yakuba* as outgroup species, from the PopDrowser database (Ramia et al. 2012). For each protein-coding gene (n_G = 14380, after excluding mitochondrial and Y-linked genes in release 5.42), we extracted the MSA of (i) its largest transcript isoform, (ii) its 4d-fold degenerate positions, (iii) its intronic regions and (iv) its upstream sequence (the 2Kb 5' upstream region, provided it does not overlap with any other protein-coding region).

Handling residual heterozigosity and missing data

The DGRP project includes a suite of 158 *D. melanogaster* lines from a natural population of Raleigh, inbred to near homozigosity. We treated residual heterozygosis as missing data. Since ~40% of the MSA positions include missing data (residual heterozigosity and/or "N"s), removing these alignment columns (complete deletion) would yield an important loss of information. Alternatively, since only ~12% of the MSA positions include eight or more missing nucleotides (Figure S1), we focused our analyses on positions with at least 150 valid alleles (n_P = 105630385 positions).

The analysis of a fixed number of 150 valid alleles along the MSA allows comparing variability among regions, and provides enough statistical power to conduct evolutionary inferences. To deal with positions including more than 150 valid alleles, we used a probabilistic approach. Suppose an alignment position with 152 valid alleles (*e.g.* 150 adenines, 2 derived timines, and 6 missing variants), from which we have to sample only 150 valid nucleotides. Three different samples could be extracted: 148 adenines and two timines, 149 adenines and one timine, or 150 adenines and zero

timines. The probability of each configuration can be calculated by means of the hypergeometric distribution. In this example, we would consider a doubleton, singleton or monomorphic mutations/position with probabilities 0.9738, 0.0261 and $8.7138e^{-05,}$ respectively.

Handling intra-locus linked positions

Most frequency spectrum-based tests assume that nucleotide positions evolve independently. However, natural selection can increase (e.g. via selective sweep) or reduce (e.g. via background selection) the frequency of nearby neutral alleles in regions with strong linkage disequilibrium (LD), such as in the *D. melanogaster* telomeric regions (Figure S2). To deal with confounding LD effects, we applied two different approaches. First, since recombination can break up LD, we only analysed the nucleotide diversity patterns at high-recombining regions (>2 cM/Mb). The recombination rates were obtained using the Recombination Rate Calculator v2.2 (Fiston-Lavier et al. 2010). Second, for each group of linked mutations (haplotype block), we only analysed the most central SNP as representative (tagSNP) of the allele frequencies within the haplotype block. For this analysis, we detected significantly linked positions by means of the Fisher exact test, after controlling for multiple testing (Benjamini and Hochberg 1995). We then cluster all these pairs into larger groups (the haplotype blocks) using mcl v10-201 program (Van Dongen 2008).

Nucleotide diversity across different site categories

We used VariScan (Vilella et al. 2005; Hutter et al. 2006) to estimate the nucleotide diversity (π) at the upstream and 4d-fold sites of each protein-coding gene (the number of protein-coding genes in high-recombining regions is n_{GHR} = 9328). In addition, we

also computed π at positions 8-30 of short introns (≤65 bp) (n_{IHR} = 12844 introns), since they may represent the most neutrally evolving class of sites (Parsch et al. 2010). For each site category, we computed the mean and confidence intervals of π over n_{GHR} (or n_{IHR}), using the wtd.mean and wtd.quantile functions of the [R] programming language.

Unfolded Site Frequency Spectrum (uSFS)

To determine the unfolded site frequency spectrum (uSFS), we reconstructed the intronic and upstream ancestral sequences of the 158 *D. melanogaster* lines. For that, we compared the consensus of such lines (selecting the most frequent nucleotide variant at each position) against the *D. simulans* and *D. yakuba* outgroup species (to polarize mutations), by means of the joint ancestral reconstruction of the 'baseml' program (Yang 2007). Given the sequence of the *D. melanogaster* ancestor, computing the uSFS (at biallelic alignment positions with at least 150 valid nucleotides) is straightforward.

Distribution of fitness effects, and rates of adaptive evolution

We used the DFE v2.03 program (Keightley and Eyre-Walker 2007; Schneider et al. 2011) to estimate the impact of natural selection at the chemosensory upstream sequences in high-recombining regions (n_{CHR} = 133). However, since the demographic events may mimic the nucleotide diversity pattern yield by natural selection, we estimated the underlying demographic history of the Raleigh population from the uSFS of the 228533 short intron positions ($uSFS_{SIP}$, used as neutral reference). In particular, we contrasted if the population has had: (i) a constant effective population size (N_e), (ii) one N_e change (expansion or contraction) some generations ago (two-epoch model), or (iii) two N_e changes (expansions or contractions) in the past (three-epoch model). The

weighted Akaike Information Criterion (wAIC) indicated that the best-fit model is the last one (three-epoch).

We then evaluated the fit of three selection regimes to the uSFS data of the upstream regions: (i) the "no selection model", which considers that the allele frequencies are just shaped by the three-epoch demographic scenario, (ii) "negative selection model", which contemplates the three-epoch demographic scenario plus the fitness effects of new deleterious mutations (DFE; modelled by means of a Gamma distribution), and (iii) the "complex selection model", which includes the three-epoch demographic scenario, the DFE, and a proportion of mutations (p_a) with certain selective advantage (s_a). To determine the impact of natural selection at the 133 chemosensory upstream regions, we compared the "no selection model" against the "negative model" (test for negative selection), and the "negative model" against the "complex model" (test for positive selection) via the likelihood ratio test (using a Chi-squared with 2 d.f.).

Functional and expression data

We used information of the *D. melanogaster* genomic regions bound by specific regulatory proteins (often represented as binding peaks, or BPs). In particular, we searched the 'TF_binding_sites', 'enhancer', 'silencer' and 'insulator' keywords in the 'feature' field of the FlyBase (Marygold et al. 2013) GFF file (release 5.54). We just retained the BPs located at the upstream regions (2Kb upstream from the translation start site) of the chemosensory gene families. To test whether the transcriptional patterns of the chemosensory genes can be explained by their BP content, we examined their expression RPKM values (Reads Per Kb and Million of mapped reads) across 80 different conditions (30 developmental stages, 29 tissues and 21 treatments) (Marygold et al. 2013). We excluded those chemosensory genes without expression changes

across conditions, because the correlation (similarity) between gene pairs cannot be computed if the standard deviation is zero.

Inference of binding motifs in *D. melanogaster*

The BP information may provide a limited resolution in the present study, since: (i) BPs are larger (~900 bp) than real CRE motifs (tipically, 6-15 bp), and include many DNA positions not involved in tuning the protein binding affinity; and (ii) each BP often represents an unknown number of CRE motifs. To overcome these limitations, we inferred the actual number of CRE motifs by scanning (Grant et al. 2011) the *D. melanogaster* chemosensory upstream sequences with the Position Weight Matrices (PWMs) stored in the JASPAR v5 database (Mathelier et al. 2013).

In the 5' chemosensory upstream regions showing statistical support for positive selection (i.e. significant for the 'positive selection' test), we also examined the polymorphism-to-divergence ratio inside and outside these inferred CREs. We did not used the DFE statistical framework, since its application may not be appropriated with a low number of positions (Schneider et al. 2011). Instead, we used the *mstatspop* program, which implements a new method especially suitable for large data sets including missing data (Ferretti et al. 2012).

Analysis of the CRE turnover

We studied the CRE turnover process occurred during the *Drosophila* diversification. For that, we first inferred the actual number of CREs motifs (*Cij*) per regulatory protein (*i*) in five closely-related species (*j*): *D. melanogaster*, D. *sechellia*, *D. simulans*, *D. yakuba* and *D. pseudoobscura*. Since the inference of each CRE motif (*t*) has a particular level of uncertainty (p_{ijt}), we weighted *Cij* as:

$$C_{ij} = int\left[\sum_{t=1}^{n_{ij}}\left(1 - p_{ijt}\right)\right]$$

where n_{ij} is the total number of inferred CRE motifs per regulatory protein (i) in species (j).

Previous molecular knowledge suggests that the evolution of C_{ij} may be consistent with a gain-and-death (GD) model, where CREs originate *de-novo* (density-independent mechanism, because the probability of having a *de novo* origin is independent on the current number of CREs), and become lost by deletion or motif degeneration (density-dependent mechanism, because the probability of having a deletion or motif degeneration depends on the actual number of CREs). In particular, we used the GD model (*-rmodel GD*) as implemented in the BadiRate program (Librado et al. 2012) to evaluate the fit of three branch scenarios to C_{ij}: (i) the GD-GR-ML model, where all lineages have the same GD rates (one branch class); (ii) a GD-dmel-ML model, where GD rates can vary between the foreground (*D. melanogaster*) and the background lineages (two branch classes); and (iii) a GD-FR-ML model, where all lineages can have particular GD rates (eight branch classes). To avoid local convergence problems, we analysed each branch model with 20 different seed values, and the one with best likelihood was selected for further analyses. The fit of the branch models to C_{ij} was evaluated via the wAIC.

Results

Binding peak data at the 5' chemosensory upstream regions

The analysis of the BP distribution in each chemosensory gene showed a highly heterogeneous distribution (Figure 1; Table S1). First, the transcription of the *Ir68a* and *Or*22a genes is controlled by at least 16 and 12 regulatory proteins, while the upstream region of the universal co-receptor ORCO only includes a single Caudal BP. Second, a few BP are gene or family-specific (e.g. Disco appears to exclusively regulate the *Obp99b*), whereas others control the transcription of multiple genes (e.g. Caudal or Chinmo). Third, in comparison with non-chemosensory genes, we found that the number of BPs is significantly lower at the OBP, OR, GR and aIR upstream regions (Wilcoxon exact test; $P < 0.05$; Table S1).

To further understand the functional significance of this particular BP landscape, we computed the transcriptional similarity among chemosensory gene pairs (Figure S3), as well as the proportion of BPs shared between their upstream regions. We remarkably observed that the BP and transcriptional similarity measures are uncorrelated (Mantel test; $P = 0.58$), suggesting that the current BP annotations are inappropriate to study the impact of natural selection in the transcriptional evolution of the chemosensory genes.

Demographic history of the *D. melanogaster* Raleigh population

Since demographic events can mimic the molecular hallmark of natural selection (Tajima 1989), we need to take into account the already-described bottleneck experienced by *D. melanogaster* during the colonization of North America (Keller 2007; Duchen et al. 2012). The molecular hallmark led by the demographic scenario can be discerned by analysing genome-wide data of a selectively neutral class of sites. We examined three candidate classes of neutral sites: the 2Kb upstream regions, the 4d-fold

degenerate sites, and the positions 8-30 of short introns (≤65 bp) (short intron positions, or SIP). In agreement with previous findings (Parsch et al. 2010), our genome-wide analysis shows that SIPs exhibit the highest levels of nucleotide diversity, with an X-to-autosomes π_{SIP} ratio that approaches to the neutral expectations of ¾ (Figure 2). Overall, it supports the idea that the SIPs are negligibly affected by natural selection.

We tested whether the unfolded site frequency spectrum of mutations at SIPs ($uSFS_{SIP}$, Figure 3) shows the footprint of a recent bottleneck event. We separately analysed the autosomic and X-linked $uSFS_{SIP}$, and consistently found that the best-fit demographic scenario is the 3-epoch model ($wAIC_x$ = 0.99, and $wAIC_{autosomic}$ = 0.99, respectively), which supports the demographic inference proposed for this population (Duchen et al. 2012). We used the ML parameter estimates (under the 3-epoch model) to control for this demographic scenario, when analysing the action of natural selection.

Impact of natural selection in the evolution of the chemosensory upstream regions

We examined the impact of natural selection by concatenating the 5' upstream regions of each chemosensory family, and applying the 'negative' and 'positive selection' tests. We remarkably found that the latter test is significant for all the chemosensory families, indicating a significant contribution of both, negative and positive selection (Table 1). The distribution of fitness effects (DFE) reveals that most new mutations are mildly ($10 < -N_e s < 100$) or strongly deleterious ($-N_e s > 100$), especially for the aIR and CSP upstream regions (Figure 4).

We also inferred huge rates of adaptive substitution (α), ranging from 0.77 in aIRs to 0.98 in ORs (Table 1). Nevertheless, since α is very sensitive –among other factors- to the demographic events (Gossmann et al. 2012), α is not directly comparable among studies, or even among genomic regions. In this regard, a better estimator of the

adaptive pressure is ω_α (α normalised by the rate of neutral substitution) (Gossmann et al. 2012). We found that the OR and GR upstream regions exhibit higher ω_α estimates than the other 5' chemosensory upstream regions (Table 1).

We also conducted the 'negative' and 'positive selection' tests separately for each chemosensory upstream gene region. After correcting for multiple testing, we notably found that 32 out of the 133 analysed upstream regions are statistically significant (Table 2). However, an in-depth analysis casts doubts on the biological relevance of 10 of them. For example, four out of these 10 upstream regions are located at the chromosome band 56, whereas five at the chromosome band 22 (Figure S4). Rather than independent gene-by-gene positive selection events, this clustering of statistically significant cases suggests some chromosome context-effects. Similarly, the *Or9a* gene is located near *nocte,* in a region that shows the molecular hallmark of a selective sweep (*nocte* is involved in the temperature compensation of the circadian clock, critical for the adaptation to temperate environments). Even after excluding these 10 uncertain cases, we remarkably found that eighteen out of the 22 remaining upstream regions belong to the OR and GR families, an overrepresentation of membrane chemoreceptor genes (Fisher test; $P = 0.0284$). This result corroborates that positive selection is more pervasive at the 5' upstream regions of these two chemoreceptor families.

Distribution of natural selection along the chemosensory upstream regions

The analysis of whole 5' upstream regions may motivate erroneous interpretations about the impact of natural selection in transcriptional evolution. Indeed, upstream regions often include many positions not involved in transcriptional regulation. To test whether the molecular hallmark of positive selection is homogeneously distributed along the 22 positively-selected upstreams regions, we compared the polymorphism-to-divergence

ratio inside (π/K_i) and outside (π/K_o) the CRE binding motifs inferred by FIMO (Figure 5). We found that π/K_i depends on the stringency level used to infer binding regions. With a FIMO qvalue cutoff equal to 0.01, π/K_i is lower than π/K_o. However, relaxing the FIMO qvalue cutoff gradually diminishes this pattern, until approaching the nearly constant π/K_o ratio found outside the inferred binding motifs. Remarkably, these π/K_i vs. π/K_o differences are very likely to result from the action of natural selection within the CRE binding sites strongly resembling its canonical motif.

Natural selection and CRE turnover

To gain insights into the impact of positive selection in driving the CRE turnover, we examined the gain and loss of binding motifs across the 5' upstream regions of each chemosensory family. In particular, we estimated the gain and death CRE turnover rates using the maximum likelihood framework provided by the BadiRate program (Librado et al. 2012). We found that the CRE gain rate ranges from $\gamma = 0.0092$ (OBPs) to 0.0152 (ORs) gains per Mya, while the CRE death rate from $\delta = 0.0086$ (aIRs) to 0.0259 (GRs) losses per Mya per CRE copy. In fact, the CRE death rate perfectly correlates with the $p_a \cdot s_a$ (the product of the rate, p_a, and strength, s_a, of advantageous, which is an excellent estimator of the impact of positive selection at the population level; Spearman rank correlation coefficient; $\rho = 1$, $P = 0$; Figure S5). On the contrary, $p_a \cdot s_a$ and γ are not significantly associated (Spearman rank correlation coefficient; $\rho = 0.1$, $P = 0.8729$). This lack of association may result from the variation in the CRE number per upstream region (Table S1), given that γ cannot be normalized by this factor (unlike δ, see methods). To test this hypothesis, we analysed the net number of turnover events, instead of the γ, δ rates. Remarkably, we inferred a higher number of gain and loss events in the 22 positively-selected than in the rest of upstream regions (Wilcoxon

test; $P = 0.0472$ for gains, and $P = 0.0316$, for losses), which strongly supports the idea that positive selection drives the gain and loss of binding motifs at the chemosensory upstream regions.

Discussion

Analyses of nucleotide diversity in large population data sets

Analyses of large population data sets open unprecedented opportunities to conduct large-scale evolutionary analyses, including the detection of the molecular hallmark of natural selection (Mackay et al. 2012; Pool et al. 2013). Large data sets, however, entail some methodological challenges, like the treatment of missing data (Ferretti et al. 2012).

Here, we used a MSA from 158 *D. melanogaster* inbred lines from a single population (Raleigh, North Caroline; DGRP project) (Mackay et al. 2012), which harbours ~40% of positions either with missing nucleotides (caused by sequencing problems) or with residual heterozygosity. Removing these alignment columns (complete deletion) will yield an important loss of information, as well as a potential bias and misleading interpretations on the evolutionary forces shaping the nucleotide diversity patterns (residual heterozygosity is unevenly distributed along the genome (Langley et al. 2012)). To handle with such problem, we applied two different strategies: to compute the uSFS using a probabilistic approach, and to compute the levels of polymorphism and divergence using the mathematical framework provided by (Ferretti et al. 2012).

The demographic history of *D. melanogaster* and natural selection

It is well know that *D. melanogaster* colonized America in recent years. In North America, the first *D. melanogaster* specimen was captured in 1875, and only 30 years later it was the most common dipteran species along the mainland (Keller 2007). The North America colonisation likely offered opportunities to natural selection, as response to the new climate environment pressures. Indeed, the colonization is so recent that the molecular footprint of positive selection may remain in the *D. melanogaster* genome

and, particularly, in some chemosensory system genes (which may have play a relevant role in the adaptive process).

The demographic events, however, can mimic the molecular hallmark of natural selection. To discern between both processes, we must analyse -as a reference- a neutral class of sites. In agreement with previous findings (Parsch 2003; Parsch et al. 2010), we found that the positions 8-30 of the short introns (≤65bp) are the most effectively neutral data, with an X-to-autosomic ratio of 0.72. Indeed, its nucleotide diversity is slightly lower than in 4d-fold degenerate positions, which may reflect a certain level of codon usage bias (CUB) for translational efficiency or accuracy (Nielsen et al. 2007; Poh et al. 2012; Lawrie et al. 2013).

The extension of the McDonald-Kreitman (MK) (McDonald and Kreitman 1991) test is not appropriated for a bottleneck scenario (Charlesworth and Eyre-Walker 2008; Parsch et al. 2009). One the one hand, a recent population expansion can cause the fixation of slightly deleterious mutations, leading to the false inference of positive selection. On the other hand, the MK test assumes that deleterious mutations are not segregating in the population, which may not hold in epochs of reduced *Ne*. In this case, slightly deleterious mutations will contribute more to polymorphism than to divergence, yielding underestimates of α. To circumvent the latter problem, Fay et al. (2001) proposed filtering out low-frequency variant positions (enriched in slightly deleterious mutations). In populations recovering from a bottleneck, nevertheless, most neutral mutations may also be segregating at low frequencies. Thus, applying the MK test may bias the results.

The huge rates of adaptive evolution

We found rates of adaptive substitution ($\alpha \approx 0.95$) much higher than those previously-inferred in the *D. simulans* non-coding regions ($\alpha \approx 0.5$) (Andolfatto 2005). In this context, it is worth noting that the mean fixation time for new mutations is:

$$t_{fix} \approx \frac{2(\log(2N_2) + \log(2s) + \gamma)}{s}$$

where N_2 is the population size after the bottleneck recovery, relative to the ancestral population size (1.7 in our case), γ is the Euler's constant (0.5772) and s_a adaptive coefficient. If $s_a \approx 0.01$, the mean fixation time is 3.94 N_2 generations, whereas only 0.06 N_2 generations are required if $s_a \approx 0.99$ (Figure S6). Therefore, the recovery from the bottleneck is so recent (0.05 N_2 generations ago) that hardly the strongly favoured mutations have had time enough to reach fixation. Consequently, our α estimates may not indicate a huge adaptive pressure to environmental changes, but rather point out the recent demographic expansion of this population (Duchen et al. 2012). To avoid misleading interpretations about the impact of positive selection, we estimated the ratio of the adaptive-to-neutral substitution rates (ω_a), an estimator that is indeed comparable among different populations and genomic regions.

Evolutionary comparison of the chemosensory upstream regions

One of the fundamental issues to understand the origin and function of the major chemosensory multigene families is to compare their evolutionary rates at different time-scales. Until this work, such studies have mainly focused on the gene turnover and on the selective constraint at the protein-coding regions (Sanchez-Gracia et al. 2009).

In agreement with these previous studies, we found two distinctive modes of gene family evolution (Figure 6). On the one hand, the aIR and CSP upstream regions mainly evolve under a strong negative selection (high $-N_e s$), with only a few mutations being

advantageous (low p_a; Table 1). It could be speculated that the low number of aIR and CSP upstream regions may compromise our ability to accurately infer the selective parameters. However, the standard errors of the estimates are comparable among the five surveyed families (Table 1), which indicates that these differences are not due to methodological issues, but they have some biological relevance. Indeed, the aIR and CSP gene families have an ancient origin (Vieira and Rozas 2011), and likely commit a basal function in sensing chemical cues (Croset et al. 2010), or even in other biological processes (Nomura et al. 1992). On the other hand, the GR and OR upstream regions exhibit the highest ω_α values (Table 1), relative to the other chemosensory gene families. This suggests that molecular impact of positive selection at the upstream region not only depends on the biological function of the families (olfaction vs. taste), but also on the molecular process in which they are involved (membrane receptor vs. extracellular binding proteins).

It is also worth noting that our estimates of positive selection ($\omega_\alpha \approx 0.4$) are comparable to that inferred for a sample of 373 *D. melanogaster* protein-coding regions (Gossmann et al. 2012). It partially contradicts some ideas postulating that upstream regions may have a large contribution to the adaptive evolution. In fact, we found that neither the rate (p_a) nor the strength (s_a) of positive selection are higher for the chemosensory than for the rest of upstream regions (Bootstrap analysis; $P > 0.05$, for all the surveyed chemosensory gene families).

These results, however, should be cautiously interpreted. Indeed, we found that the molecular hallmark of natural selection is not homogeneously distributed along the upstream regions, but differentially focalised within the CREs (Figure 5). Since the number of CREs varies for each chemosensory upstream region, the analysis of the whole upstream region may yield misleading interpretations.

In this regard, it has been found that the CRE content highly varies across closely-related species (Bradley et al. 2010; Dowell 2010; Schmidt et al. 2010). This huge turnover has been traditionally interpreted as a result of genetic drift and compensatory gains and losses (Ludwig et al. 2000; Durrett and Schmidt 2008). However, and consistently with recent findings (He et al. 2011), we detected that the $p_a \cdot s_a$ product is strongly correlated with the CRE turnover, supporting the hypothesis that positive selection drives the gain and loss of CRE binding motifs at the chemosensory upstream regions.

References

Andolfatto P. 2005. Adaptive evolution of non-coding DNA in Drosophila. Nature 437:1149–1152.

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

Bradley RK, Li XY, Trapnell C, Davidson S, Pachter L, Chu HC, Tonkin LA, Biggin MD, Eisen MB. 2010. Binding site turnover produces pervasive quantitative changes in transcription factor binding between closely related Drosophila species. PLoS Biol 8:e1000343.

Charlesworth J, Eyre-Walker A. 2008. The McDonald-Kreitman test and slightly deleterious mutations. Mol Biol Evol 25:1007–1015.

Clark AG, Eisen MB, Smith DR, et al. 2007. Evolution of genes and genomes on the Drosophila phylogeny. Nature 450:203 – 218.

Croset V, Rytz R, Cummins SF, Budd A, Brawand D, Kaessmann H, Gibson TJ, Benton R. 2010. Ancient Protostome Origin of Chemosensory Ionotropic Glutamate Receptors and the Evolution of Insect Taste and Olfaction. PLoS Genet 6:e1001064.

Van Dongen S. 2008. Graph Clustering Via a Discrete Uncoupling Process. SIAM J Matrix Anal Appl 30:121–141.

Dowell RD. 2010. Transcription factor binding variation in the evolution of gene regulation. Trends Genet 26:468–475.

Duchen P, Zivkovic D, Hutter S, Stephan W, Laurent S. 2012. Demographic inference reveals African and European admixture in the North American Drosophila melanogaster population. Genetics 193:291–301.

Durrett R, Schmidt D. 2008. Waiting for two mutations: with applications to regulatory sequence evolution and the limits of Darwinian evolution. Genetics 180:1501–1509.

Egea R, Casillas S, Barbadilla A. 2008. Standard and generalized McDonald-Kreitman test: a website to detect selection by comparing different classes of DNA sites. Nucleic Acids Res 36:W157–62.

Eyre-Walker A, Keightley PD. 2009. Estimating the rate of adaptive molecular evolution in the presence of slightly deleterious mutations and population size change. Mol Biol Evol 26:2097–2108.

Fay JC, Wyckoff GJ, Wu CI. 2001. Positive and negative selection on the human genome. Genetics 158:1227–1234.

Ferretti L, Raineri E, Ramos-Onsins S. 2012. Neutrality tests for sequences with missing data. Genetics 191:1397–1401.

Fiston-Lavier AS, Singh ND, Lipatov M, Petrov DA. 2010. Drosophila melanogaster recombination rate calculator. Gene 463:18–20.

Foret S, Maleszka R. 2006. Function and evolution of a gene family encoding odorant binding-like proteins in a social insect, the honey bee (Apis mellifera). Genome Res 16:1404 – 1413.

Gossmann TI, Keightley PD, Eyre-Walker A. 2012. The effect of variation in the effective population size on the rate of adaptive molecular evolution in eukaryotes. Genome Biol Evol 4:658–667.

Grant CE, Bailey TL, Noble WS. 2011. FIMO: scanning for occurrences of a given motif. Bioinformatics 27:1017–1018.

Haddrill PR, Bachtrog D, Andolfatto P. 2008. Positive and negative selection on noncoding DNA in Drosophila simulans. Mol Biol Evol 25:1825–1834.

He BZ, Holloway AK, Maerkl SJ, Kreitman M. 2011. Does positive selection drive transcription factor binding site turnover? A test with Drosophila cis-regulatory modules. PLoS Genet 7:e1002053.

Hildebrand JG, Shepherd GM. 1997. Mechanisms of olfactory discrimination: converging evidence for common principles across phyla. Annu. Rev. Neurosci. 20:595–631.

Hutter S, Vilella AJ, Rozas J. 2006. Genome-wide DNA polymorphism analyses using VariScan. BMC Bioinformatics 7:409.

Jenkins DL, Ortori CA, Brookfield JF. 1995. A test for adaptive change in DNA sequences controlling transcription. Proc Biol Sci 261:203–207.

Jensen JD, Bachtrog D. 2011. Characterizing the influence of effective population size on the rate of adaptation: Gillespie's Darwin domain. Genome Biol Evol 3:687–701.

Keightley PD, Eyre-Walker A. 2007. Joint inference of the distribution of fitness effects of deleterious mutations and population demography based on nucleotide polymorphism frequencies. Genetics 177:2251–2261.

Keightley PD, Lercher MJ, Eyre-Walker A. 2005. Evidence for widespread degradation of gene control regions in hominid genomes. PLoS Biol 3:e42.

Keller A. 2007. Drosophila melanogaster's history as a human commensal. Curr. Biol. CB 17:R77–81.

Kopp A, Barmina O, Hamilton AM, Higgins L, McIntyre LM, Jones CD. 2008. Evolution of gene expression in the Drosophila olfactory system. Mol Biol Evol 25:1081–1092.

Krieger MJ, Ross KG. 2002. Identification of a major gene regulating complex social behavior. Science 295:328 – 332.

Langley CH, Stevens K, Cardeno C, et al. 2012. Genomic variation in natural populations of Drosophila melanogaster. Genetics 192:533–598.

Lawrie DS, Messer PW, Hershberg R, Petrov DA. 2013. Strong purifying selection at synonymous sites in D. melanogaster. PLoS Genet. 9:e1003527.

Librado P, Rozas J. 2013. Uncovering the functional constraints underlying the genomic organization of the odorant-binding protein genes. Genome Biol. Evol. 5:2096–2108.

Librado P, Vieira FG, Rozas J. 2012. BadiRate: estimating family turnover rates by likelihood-based methods. Bioinformatics 28:279–281.

Ludwig MZ, Bergman C, Patel NH, Kreitman M. 2000. Evidence for stabilizing selection in a eukaryotic enhancer element. Nature 403:564–567.

Mackay TF, Richards S, Stone EA, et al. 2012. The Drosophila melanogaster Genetic Reference Panel. Nature 482:173–178.

Marygold SJ, Leyland PC, Seal RL, Goodman JL, Thurmond J, Strelets VB, Wilson RJ, the FlyBase consortium. 2013. FlyBase: improvements to the bibliography. Nucleic Acids Res. 41:D751–DD757.

Mathelier A, Zhao X, Zhang AW, et al. 2013. JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. Nucleic Acids Res. 42:D142–D147.

Matsuo T, Sugaya S, Yasukawa J, Aigaki T, Fuyama Y. 2007. Odorant-binding proteins OBP57d and OBP57e affect taste perception and host-plant preference in Drosophila sechellia. PLoS Biol 5:e118.

McBride CS, Arguello JR, O'Meara BC. 2007. Five Drosophila genomes reveal nonneutral evolution and the signature of host specialization in the chemoreceptor superfamily. Genetics 177:1395–1416.

McDonald JH, Kreitman M. 1991. Adaptive protein evolution at the Adh locus in Drosophila. Nature 351:652–654.

Nielsen R, Bauer DuMont VL, Hubisz MJ, Aquadro CF. 2007. Maximum likelihood estimation of ancestral codon usage bias parameters in Drosophila. Mol Biol Evol 24:228–235.

Nomura A, Kawasaki K, Kubo T, Natori S. 1992. Purification and localization of p10, a novel protein that increases in nymphal regenerating legs of Periplaneta americana (American cockroach). Int. J. Dev. Biol. 36:391–398.

Parsch J, Novozhilov S, Saminadin-Peter SS, Wong KM, Andolfatto P. 2010. On the utility of short intron sequences as a reference for the detection of positive and negative selection in Drosophila. Mol Biol Evol 27:1226–1234.

Parsch J, Zhang Z, Baines JF. 2009. The influence of demography and weak selection on the McDonald-Kreitman test: an empirical study in Drosophila. Mol Biol Evol 26:691–698.

Parsch J. 2003. Selective constraints on intron evolution in Drosophila. Genetics 165:1843–1851.

Pelosi P. 1996. Perireceptor events in olfaction. J Neurobiol 30:3 – 19.

Poh YP, Ting CT, Fu HW, Langley CH, Begun DJ. 2012. Population genomic analysis of base composition evolution in Drosophila melanogaster. Genome Biol Evol 4:1245–1255.

Pool JE, Corbett-Detig RB, Sugino RP, et al. 2013. Population Genomics of sub-saharan Drosophila melanogaster: African diversity and non-African admixture. PLoS Genet 8:e1003080.

Ramia M, Librado P, Casillas S, Rozas J, Barbadilla A. 2012. PopDrowser: the Population Drosophila Browser. Bioinformatics 28:595–596.

Rollmann SM, Wang P, Date P, West SA, Mackay TF, Anholt RR. 2010. Odorant receptor polymorphisms and natural variation in olfactory behavior in Drosophila melanogaster. Genetics 186:687–697.

Roy S, Ernst J, Kharchenko PV, et al. 2010. Identification of functional elements and regulatory circuits by Drosophila modENCODE. Science 330:1787–1797.

Sánchez-Gracia A, Rozas J. 2008. Divergent evolution and molecular adaptation in the Drosophila odorant-binding protein family: inferences from sequence variation at the OS-E and OS-F genes. BMC Evol. Biol. 8:323.

Sanchez-Gracia A, Vieira FG, Rozas J. 2009. Molecular evolution of the major chemosensory gene families in insects. Heredity 103:208–216.

Schmidt D, Wilson MD, Ballester B, et al. 2010. Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. Science 328:1036–1040.

Schneider A, Charlesworth B, Eyre-Walker A, Keightley PD. 2011. A method for inferring the rate of occurrence and fitness effects of advantageous mutations. Genetics 189:1427–1437.

Smadja C, Butlin RK. 2009. On the scent of speciation: the chemosensory system and its role in premating isolation. Heredity 102:77–97.

Stortkuhl KF, Kettler R, Fischer S, Hovemann BT. 2005. An increased receptive field of olfactory receptor Or43a in the antennal lobe of Drosophila reduces benzaldehyde-driven avoidance behavior. Chem Senses 30:81–87.

Tajima F. 1989. The effect of change in population size on DNA polymorphism. Genetics 123:597–601.

Vieira FG, Rozas J. 2011. Comparative Genomics of the Odorant-Binding and Chemosensory Protein Gene Families across the Arthropoda: Origin and evolutionary history of the chemosensory system. Genome Biol Evol 3.

Vilella AJ, Blanco-Garcia A, Hutter S, Rozas J. 2005. VariScan: Analysis of evolutionary patterns from large-scale DNA sequence polymorphism data. Bioinformatics 21:2791–2793.

Wang P, Lyman RF, Mackay TF, Anholt RR. 2010. Natural variation in odorant recognition among odorant-binding proteins in Drosophila melanogaster. Genetics 184:759–767.

Whiteman NK, Pierce NE. 2008. Delicious poison: genetics of Drosophila host plant preference. Trends Ecol. Evol. 23:473–478.

Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol 24:1586–1591.

Figure Legends

Figure 1. Network representation of the binding peaks (squares) bound at the 5' chemosensory upstream regions (diamonds). Diamond colours depict the chemosensory gene families. Chemosensory upstream regions without any binding peak are not represented.

Figure 2. Nucleotide diversity levels (π) at three putatively-neutral classes of sites: the 4d-fold degenerate sites, the positions 8-30 of short introns (≤65 bp), and 2kb upstream regions. π was separately computed for all the autosomic and X-linked regions.

Figure 3. The unfolded site frequency spectrum for biallelic sites, calculated from the positions 8-30 of all short introns (≤65 bp) across the autosomic and X chromosomes.

Figure 4. Distribution of fitness effects of new deleterious mutations, discretized into four categories: nearly neutral ($0<-N_e s<1$), slightly deleterious ($1<-N_e s<10$), mildly deleterious ($10<-N_e s<100$) and strongly deleterious ($100<-N_e s$).

Figure 5. The π/K ratio inside (blue line) and outside (dashed violet line) the CRE binding motifs, inferred using different FIMO qvalue cutoffs.

Figure 6. Impact of natural selection at the major chemosensory gene families. Gene turnover rate (λ), selective constraint (ω) for their protein-coding regions. Mean deleterious effect ($-N_e s$), rate of adaptive-to-neutral substitution (ω_a), proportion of adaptive mutations (p_a) and their strength (p_a). For displaying purposes, λ was multiplied by 10, while $-N_e s$ divided by 100.

Supplementary material

Figure S1. Missing allele distribution along the 2R chromosome of the 158 *D. melanogaster* lines (DGRP project).

Figure S2. PopDrowser screenshot showing polymorphism and divergence along the 2L chromosome of the 158 *D. melanogaster* lines (DGRP project).

Figure S3. Heatmap showing the transcriptional correlation values between chemosensory gene pairs. The colour gradient indicates the intensity of the correlation, from r = -1 (blue) to r = 1 (red). The gene names with an '*' are significant for the 'positive selection' test, while those with a '+' have *D. melanogaster* lineage-specific CRE turnover rates. Genes without transcriptional changes across the examined expression conditions are not included.

Figure S4. PopDrowser screenshot showing the *Gr22e-a* gene cluster region, altogether with the proportion of missing data, the nucleotide diversity, the linkage disequilibrium, and the divergence to *D. yakuba* and *D. simulans*.

Figure S5. Correlation between the CRE death rate (δ) and the product between the rate (p_a) and strength (s_a) of advantageous mutations across the 5 major chemosensory families.

Figure S6. Relationship between the mean fixation time (measured in generations/N_e) of new mutations and the coefficient of selection (s_a).

Table 1. Impact of natural selection in the concatenated chemosensory upstream sequences.

	G	AG	AP	P-value	$-N_e s$	(se)	ω_a (se)
CSP	4	3	3821	$2.30e^{-07}$	119.02 ($3.54e^{-03}$)	0.93 ($4.03e^{-04}$)	0.32 ($4.33e^{-04}$)
OBP	52	37	36449	$1.57e^{-49}$	29.68 ($3.01e^{-03}$)	0.94 ($3.11e^{-04}$)	0.35 ($1.91e^{-04}$)
OR	60	36	41575	$7.57e^{-84}$	27.77 (0.14)	0.98 ($1.58e^{-04}$)	0.42 ($1.53e^{-04}$)
GR	63	51	41666	$8.75e^{-81}$	26.05 (0.05)	0.97 ($2.89e^{-04}$)	0.44 ($1.92e^{-04}$)
aIR	15	6	6684	$1.43e^{-06}$	43.91 (0.22)	0.77 ($1.22e^{-03}$)	0.27 ($5.67e^{-04}$)

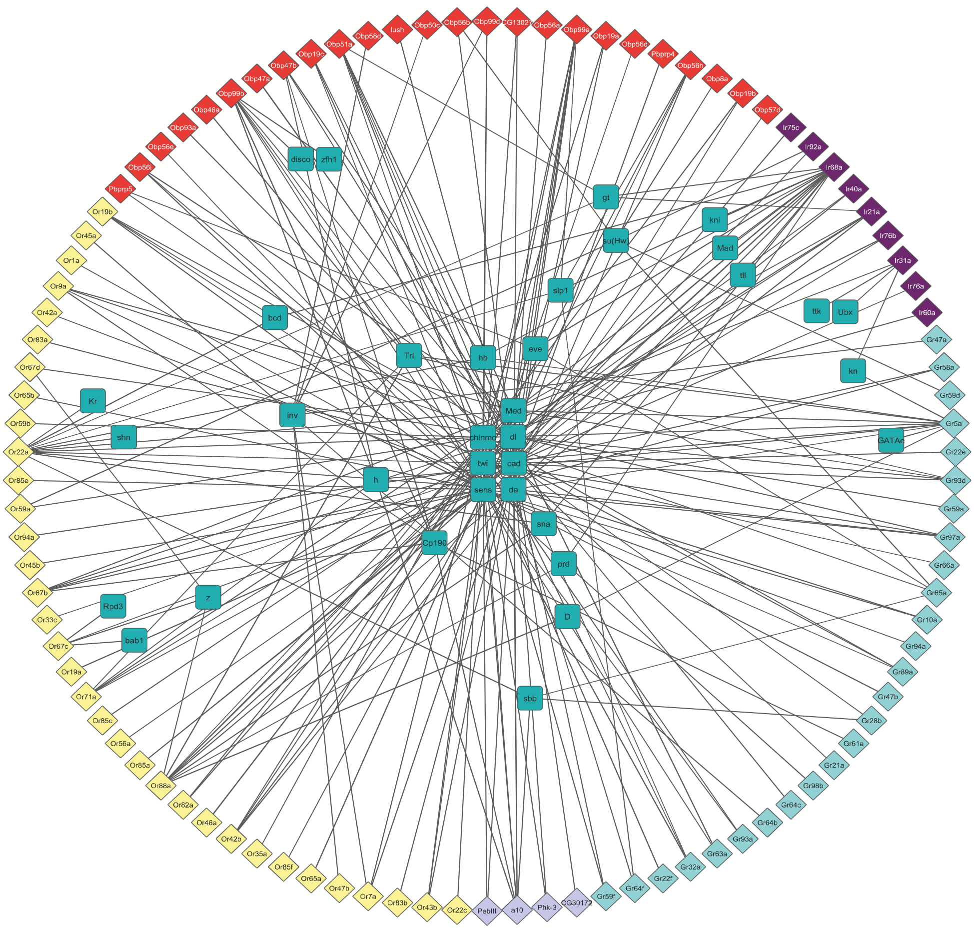
The 'G' indicates the number of gene family members, while 'AG' and 'AP' the number of genes and positions analysed with DFE v2.03. The 'P-value' column shows the p-value of the 'positive selection' test (LRT, using a Chi-Squared distribution with 2 d.f.). As the 'complex' is the best-fit model in all cases (includes negative and positive selection), the '$-N_e s_{(se)}$', '$_{(se)}$' and '$\omega_{a(se)}$' columns report the mean deleterious effect, the rate of adaptive substitutions ( ), and the adaptive-to-neutral substitution rates (ω_a), with their corresponding standard errors.

Table 2. List of the 32 upstream regions significant for test of 'positive selection'

	L(negative)	L(directional)	P-value
Phk-3	-319.10	-306.61	$3.77e^{-06}$
Obp56e*	-421.46	-411.66	$5.54e^{-05}$
Obp56d*	-416.30	-406.68	$6.61e^{-05}$
Obp46a	-451.26	-442.04	$9.95e^{-05}$
Obp51a	-469.80	-461.50	$2.48e^{-04}$
Obp99c	-446.03	-438.65	$6.23e^{-04}$
Obp56c*	-399.50	-392.56	$9.68e^{-04}$
Or22a*	-521.86	-503.93	$1.63e^{-08}$
Or69a	-425.18	-411.97	$1.83e^{-06}$
Or49a	-353.62	-340.98	$3.25e^{-06}$
Or65b	-354.56	-343.02	$9.73e^{-06}$
Or98a	-343.72	-332.68	$1.61e^{-05}$
Or22c*	-410.45	-399.74	$2.23e^{-05}$
Or67c	-406.30	-396.81	$7.57e^{-05}$
Or56a*	-305.05	-296.34	$1.64e^{-04}$
Or33a	-443.63	-435.41	$2.68e^{-04}$
Or65a	-455.91	-447.81	$3.03e^{-04}$
Or63a	-230.42	-222.47	$3.52e^{-04}$
Or9a*	-193.25	-185.42	$3.94e^{-04}$
Or59a	-523.98	-516.89	$8.34e^{-04}$
Gr68a	-511.47	-493.25	$1.23e^{-08}$
Gr59d	-485.80	-468.60	$3.37e^{-08}$
CG32395	-376.99	-362.42	$4.71e^{-07}$
Gr93d	-467.63	-454.09	$1.32e^{-06}$
Gr22e*	-489.19	-475.90	$1.69e^{-06}$
Gr22a*	-399.02	-388.05	$1.72e^{-05}$
Gr98b	-378.04	-367.46	$2.54e^{-05}$
Gr22f*	-482.38	-471.94	$2.90e^{-05}$
Gr5a	-378.50	-368.09	$3.03e^{-05}$
Gr93a	-394.65	-384.75	$5.00e^{-05}$
Gr92a	-420.14	-411.70	$2.16e^{-04}$
Gr98a	-360.26	-353.12	$7.92e^{-04}$

The 'L(negative)' and 'L(positive)' columns indicate the likelihoods of the negative and complex selection models, respectively. The asterisks denote the 10 cases excluded. The 'P-value' column shows the p-value of the 'positive selection' test (LRT, using a Chi-Squared distribution with 2 d.f.).

Figure 1



Pbprp5
Obp56i
Obp56e
Obp93a
Obp46a
Obp99b
Obp47a
Obp47b
Obp19c
Obp51a
Obp58d
lush
Obp50c
Obp56b
Obp99d
CG1302
Obp56a
Obp99c
Obp19a
Obp56d
Pbprp4
Obp56h
Obp8a
Obp19b
Obp57d
Ir75c
Ir92a
Ir68a
Ir40a
Ir21a
Ir76b
Ir31a
Ir76a
Ir60a
Gr47a
Gr58a
Gr59d
Gr5a
Gr22e
Gr93d
Gr59a
Gr97a
Gr66a
Gr65a
Gr10a
Gr94a
Gr89a
Gr47b
Gr28b
Gr61a
Gr21a
Gr98b
Gr64c
Gr64b
Gr93a
Gr63a
Gr32a
Gr22f
Gr64f
Gr59f
CG30172
Phk-3
a10
PebIII
Or22c
Or43b
Or83b
Or7a
Or47b
Or65a
Or85f
Or35a
Or42b
Or46a
Or82a
Or88a
Or85a
Or56a
Or85c
Or71a
Or19a
Or67c
Or33c
Or67b
Or45b
Or94a
Or59a
Or85e
Or22a
Or59b
Or65b
Or67d
Or83a
Or42a
Or9a
Or1a
Or45a
Or19b
disco
zfh1
gt
su(Hw
kni
Mad
tll
slp1
ttk
Ubx
bcd
Trl
hb
eve
kn
Kr
inv
Med
shn
chinmo
dl
GATAe
twi
cad
h
sens
da
sna
Cp190
prd
Rpd3
z
D
bab1
sbb

Figure 2

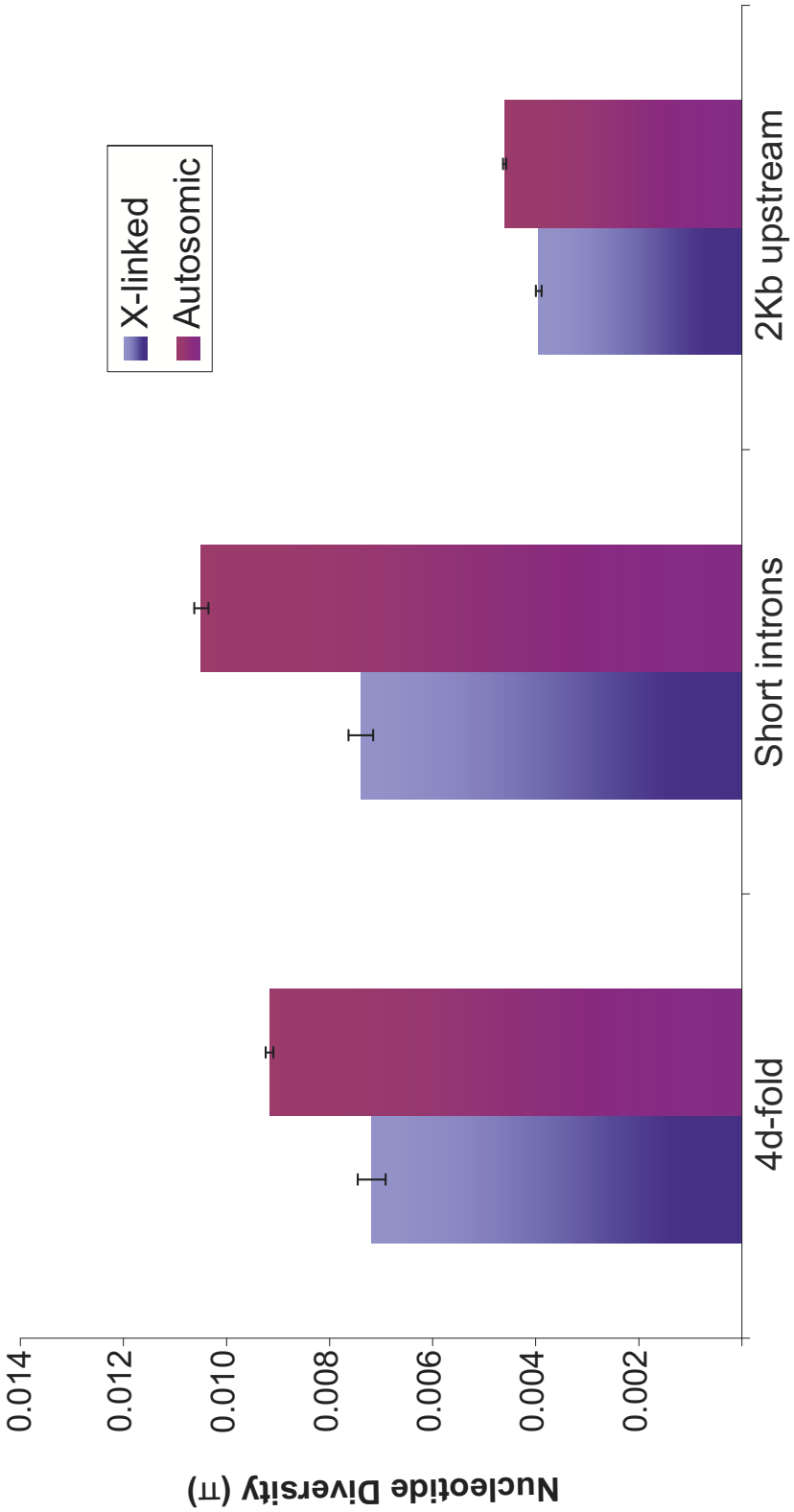

Nucleotide Diversity (π)
0.014
0.012
0.010
0.008
0.006
0.004
0.002
X-linked
Autosomic
4d-fold
Short introns
2Kb upstream

Figure 3

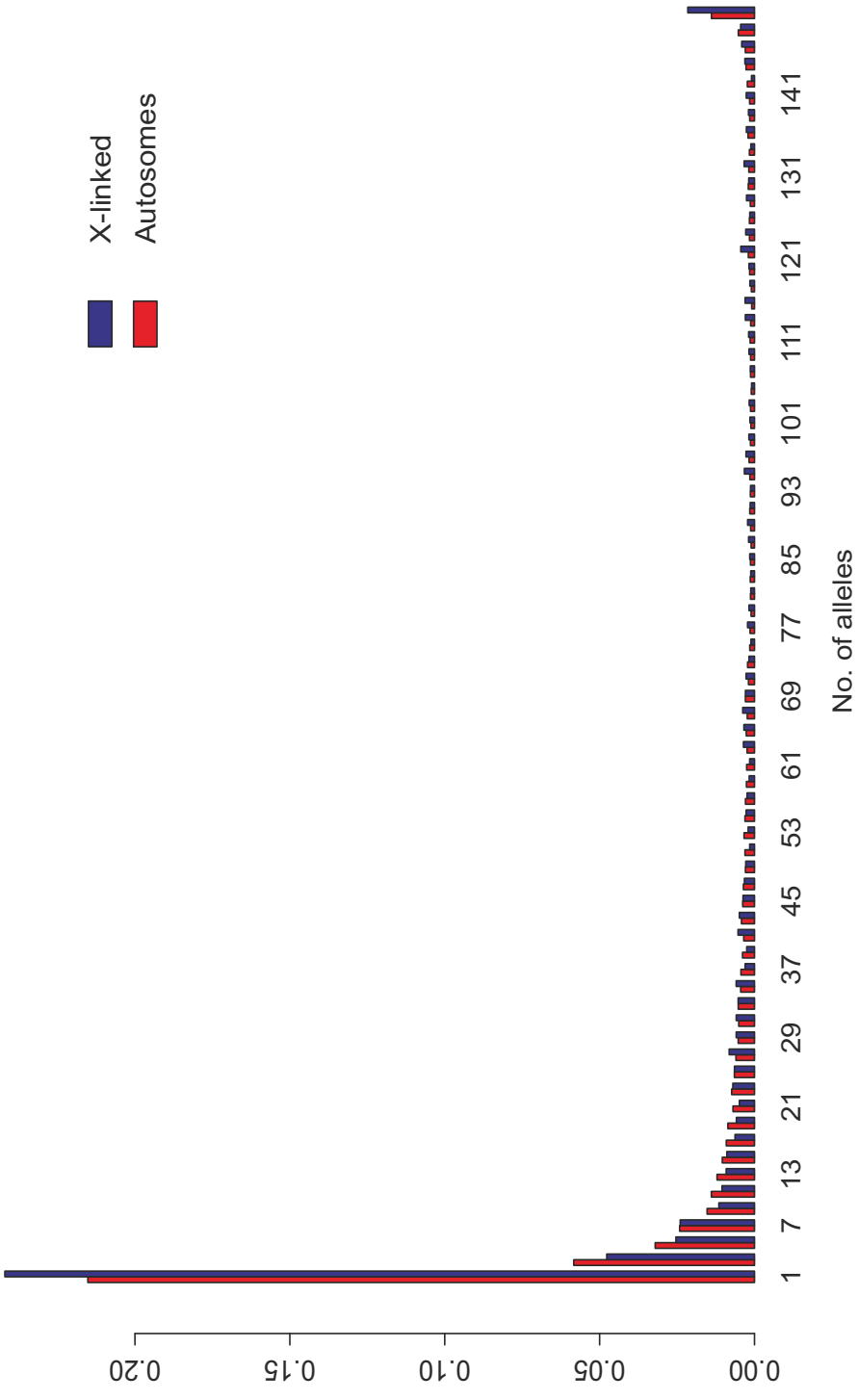

X-linked
Autosomes
0.00
0.05
0.10
0.15
0.20
1
7
13
21
29
37
45
53
61
69
77
85
93
101
111
121
131
141
No. of alleles

Figure 4

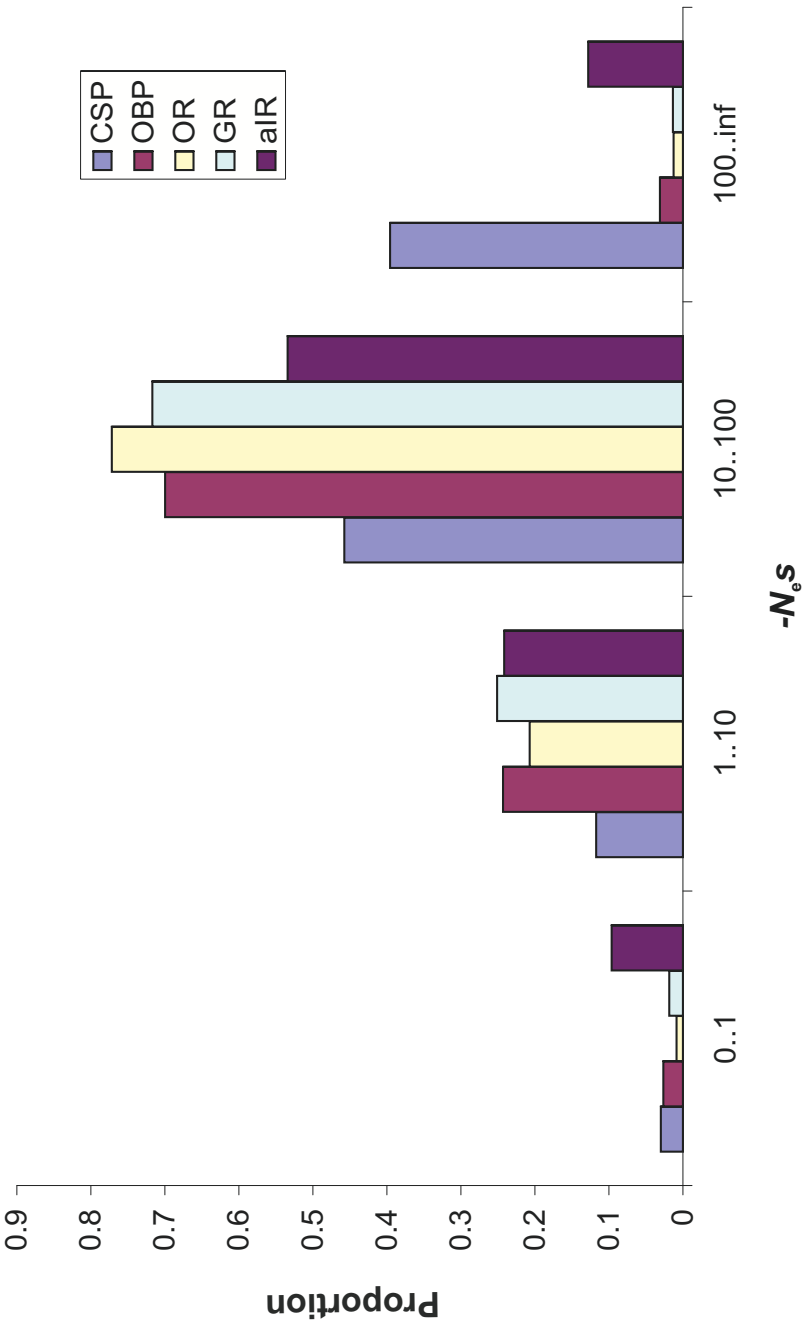

CSP
OBP
OR
GR
alR
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
Proportion
0..1
1..10
10..100
100..inf
-N_e s

Figure 5

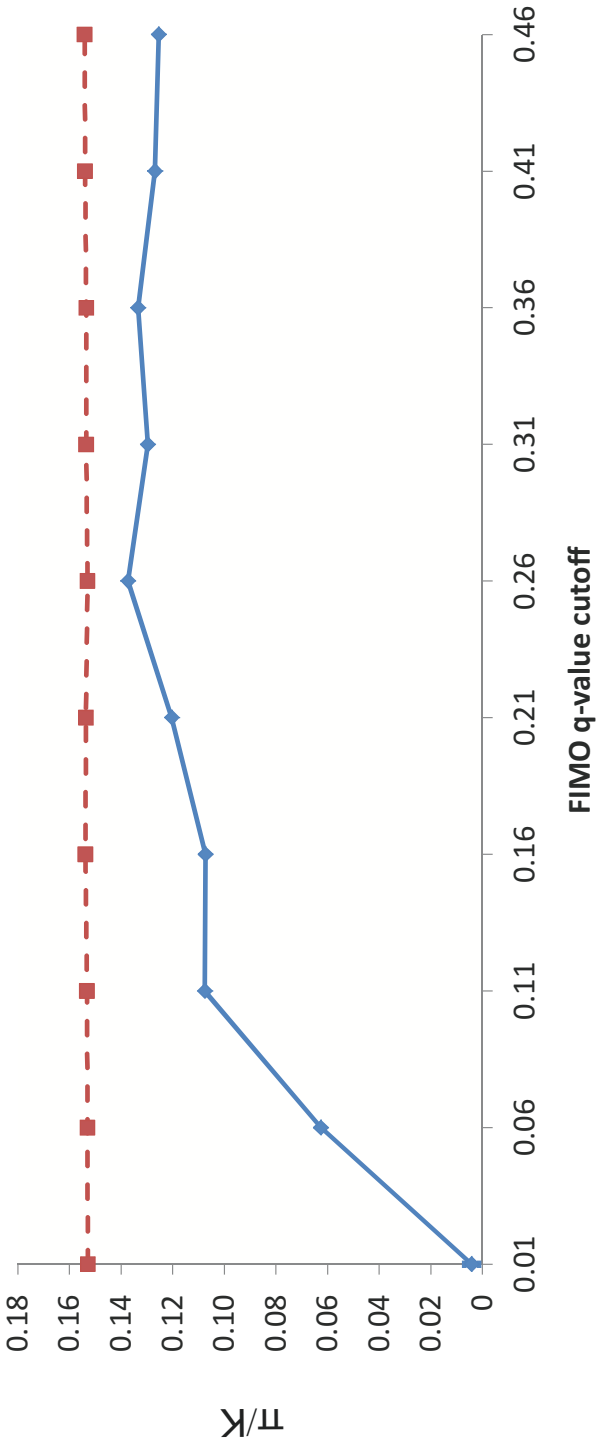
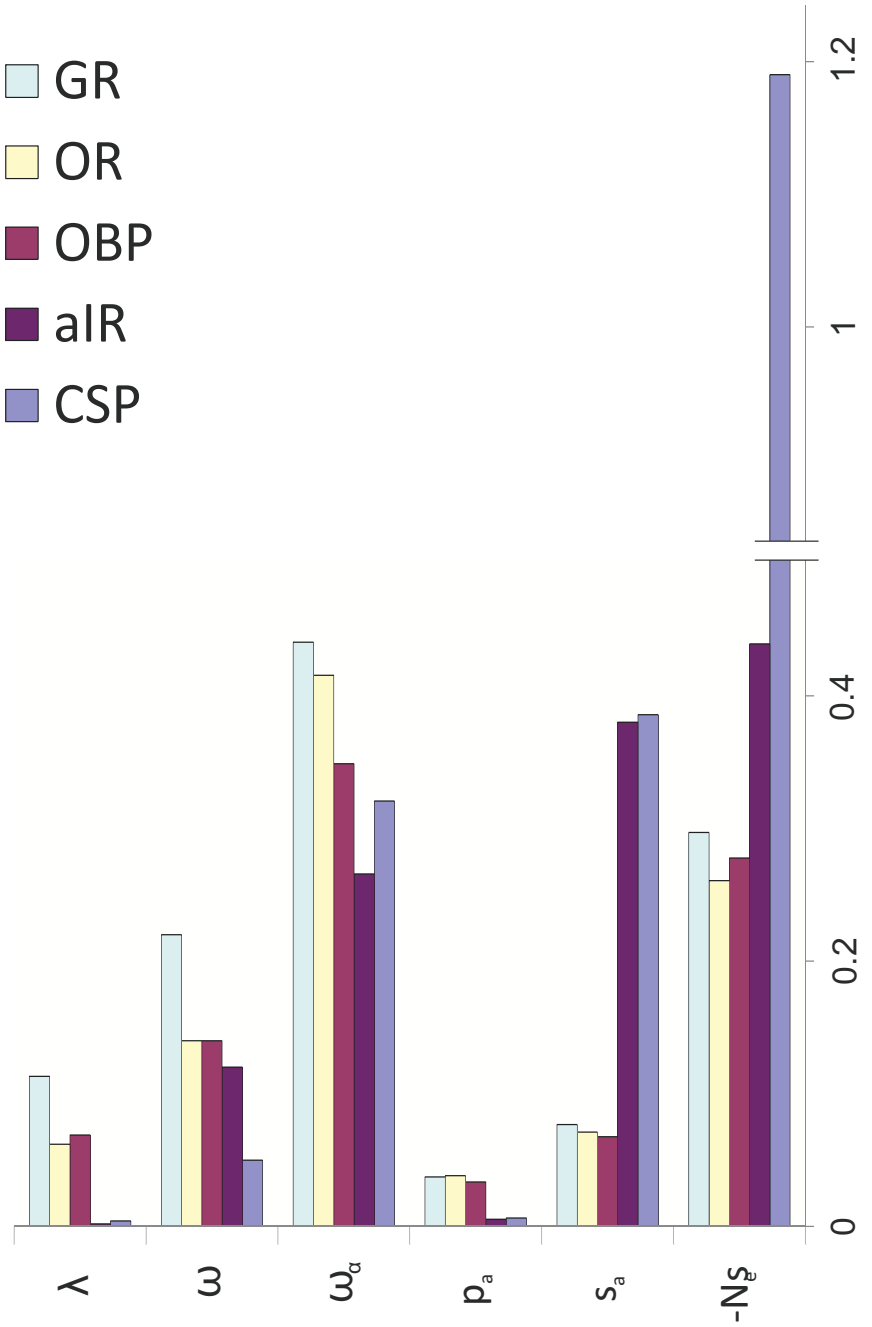

π/K
0.18
0.16
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0
0.01
0.06
0.11
0.16
0.21
0.26
0.31
0.36
0.41
0.46
FIMO q-value cutoff

Figure 6



Supplementary Material

Positive selection drives the evolution of transcriptional regulatory upstream regions of the major chemosensory gene families

Pablo Librado and Julio Rozas

Departament de Genètica & Institut de Recerca de la Biodiversitat (IRBio),
Universitat de Barcelona, Barcelona, Spain

Table S1. Distribution of BP per gene family

	Mean	Median	SE	P-value
Genome	3.86	2	0.0388	NA
CSP	1.25	1.5	0.4787	0.4206
OBP	0.83	0	0.2086	$6.506e^{-08}$
OR	1.20	1	0.2820	$4.041e^{-06}$
GR	0.77	0	0.1964	$2.868e^{-10}$
alR	1.81	0.5	1.0009	0.02998

The 'Mean', 'Median' and 'SE' (standard error) of number of binding peaks (BPs) per chemosensory upstream region. The 'P-value' column shows the probability that the number of BPs differs from the observed in non-chemosensory genes.

Figure S1

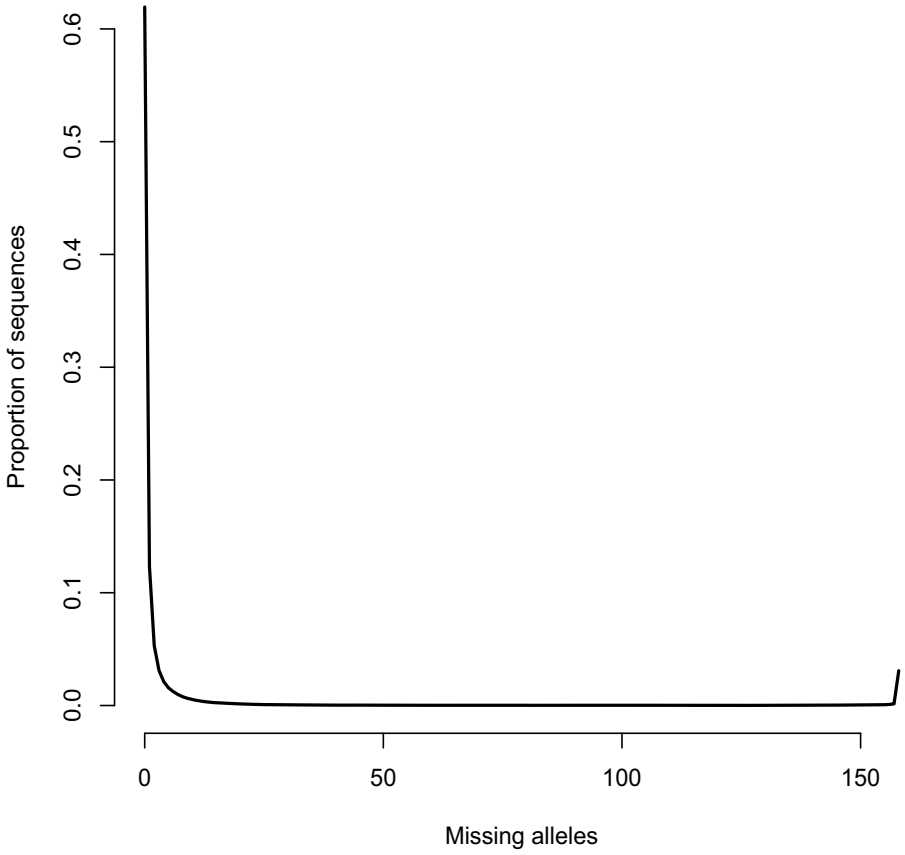

Proportion of sequences
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0
50
100
150
Missing alleles

Figure S2

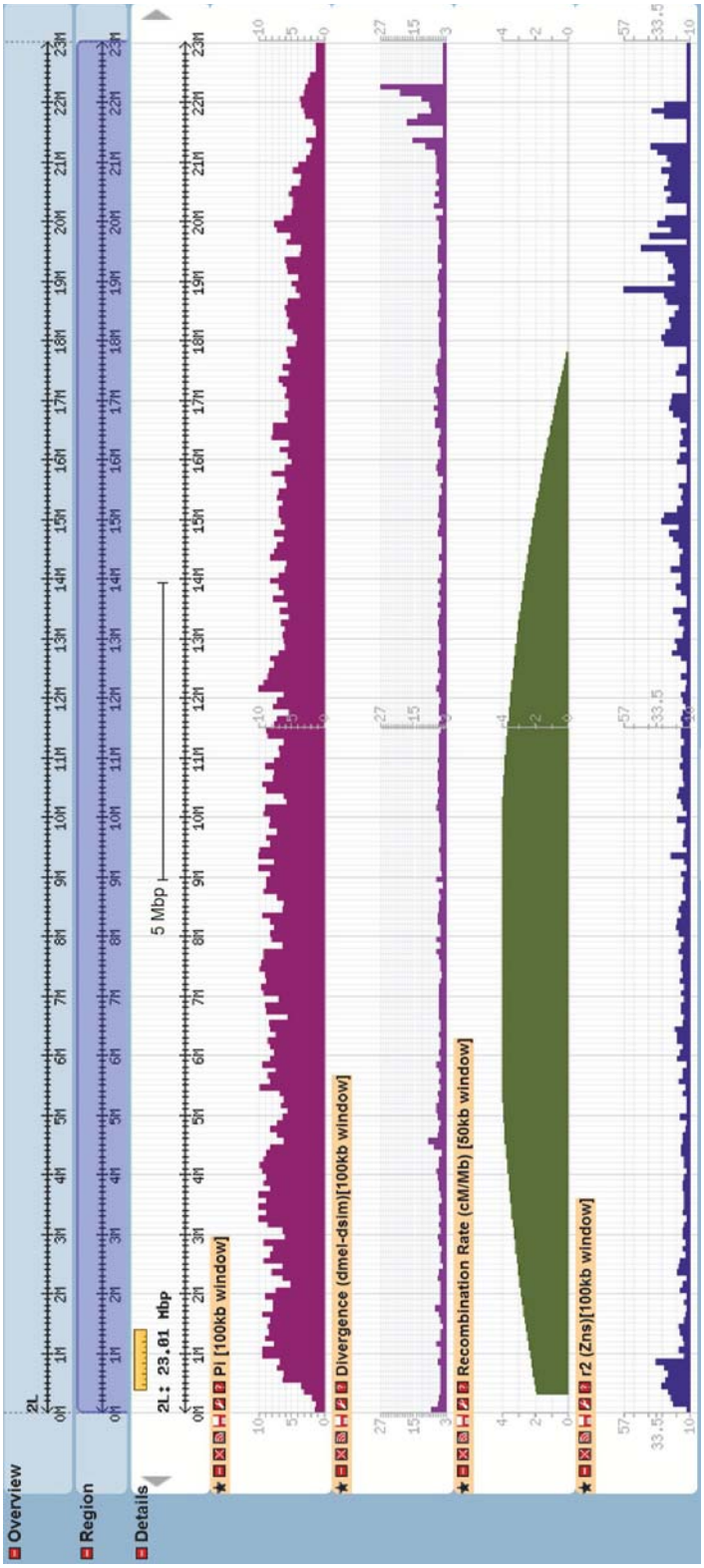

Overview
Region
Details
2L
2L: 23,01 Mbp
5 Mbp
Pi [100kb window]
Divergence (dmel-dsim)[100kb window]
Recombination Rate (cM/Mb) [50kb window]
r2 (Zns)[100kb window]

Figure S3

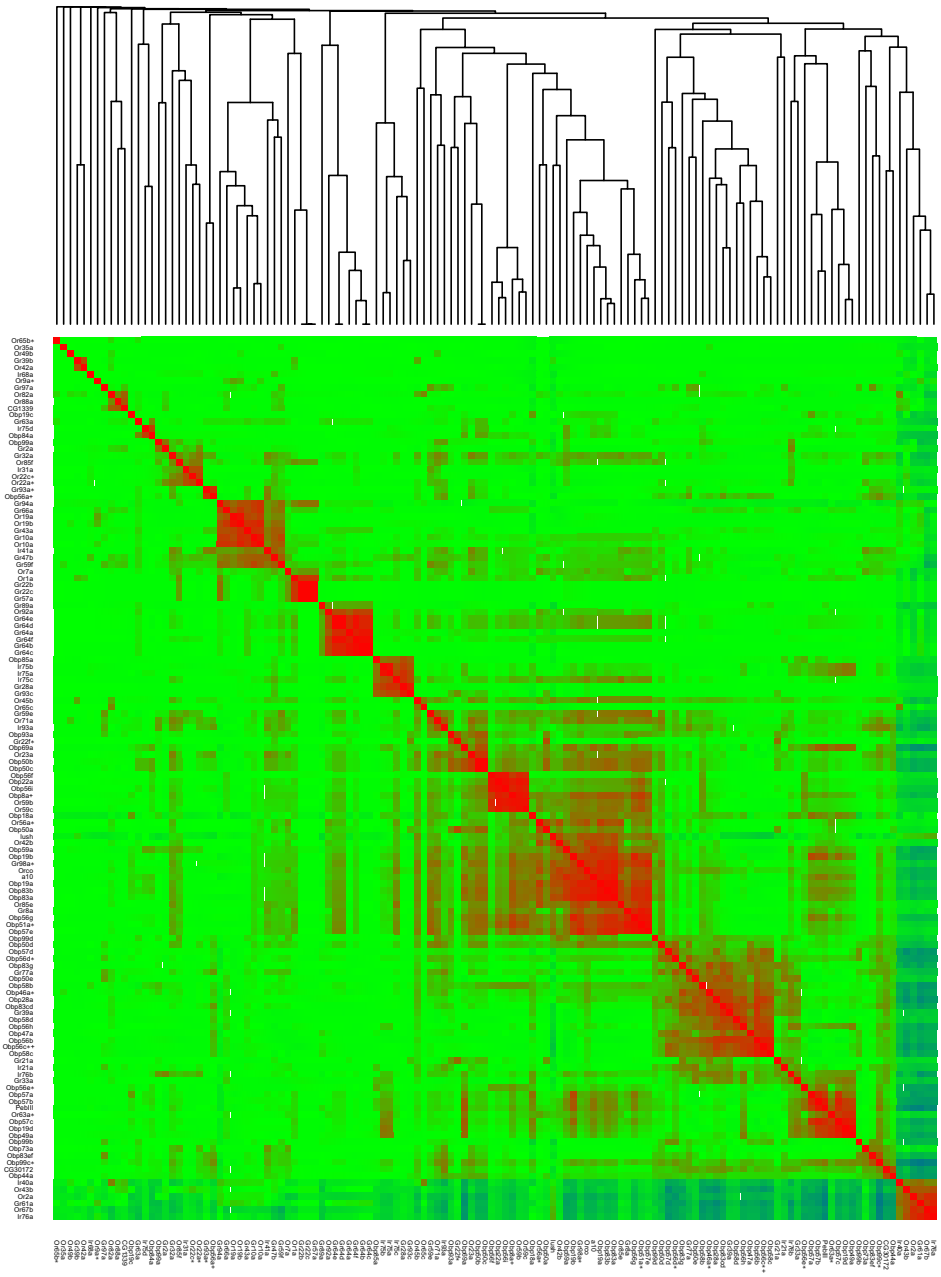


Figure S4

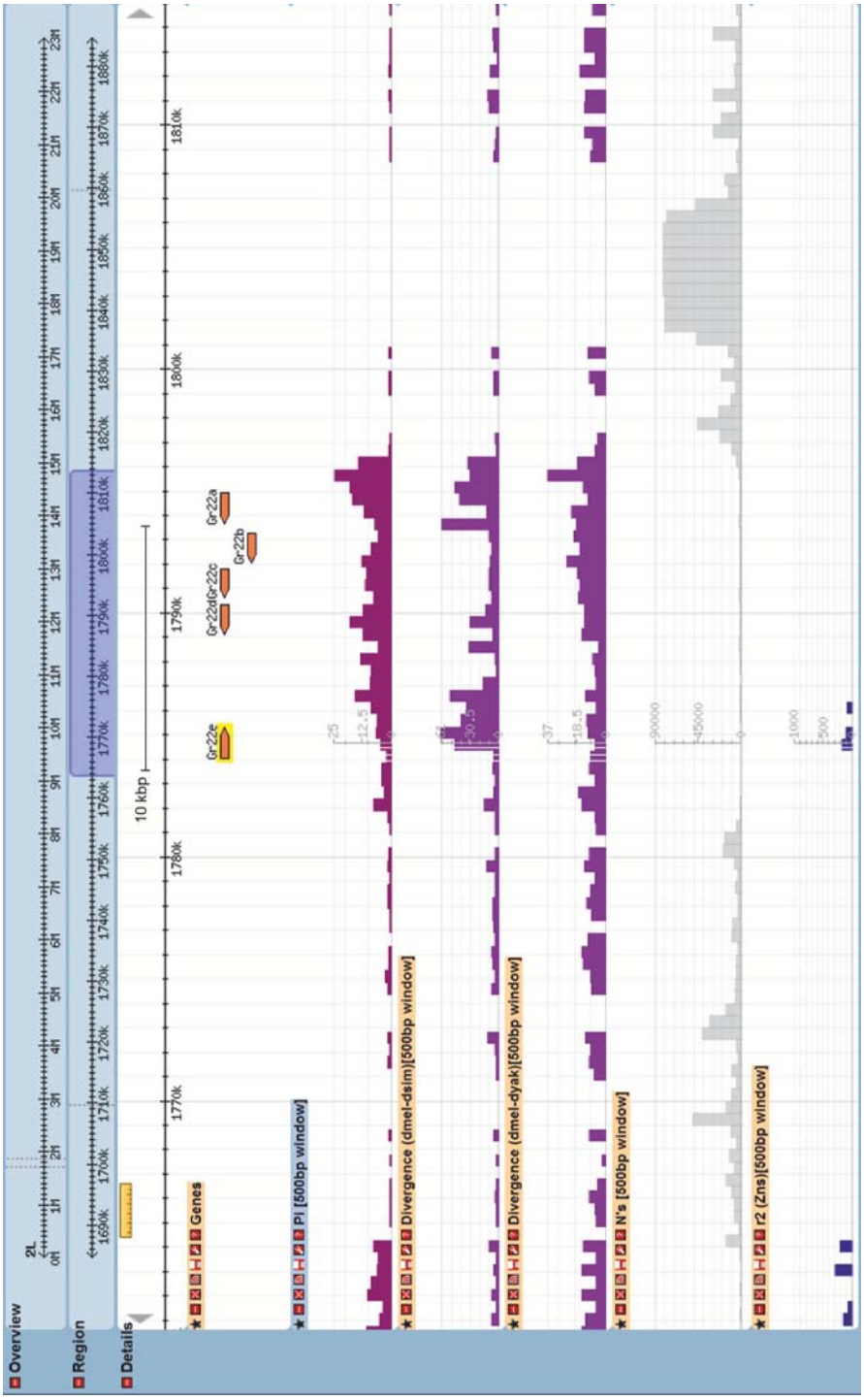

Overview
2L
Region
Details
10 kbp
Gr-22e
Gr-22d
Gr-22c
Gr-22b
Gr-22a
Genes
Pi [500bp window]
Divergence (dmel-dsim)[500bp window]
Divergence (dmel-dyak)[500bp window]
N's [500bp window]
r2 (Zns)[500bp window]

Figure S5

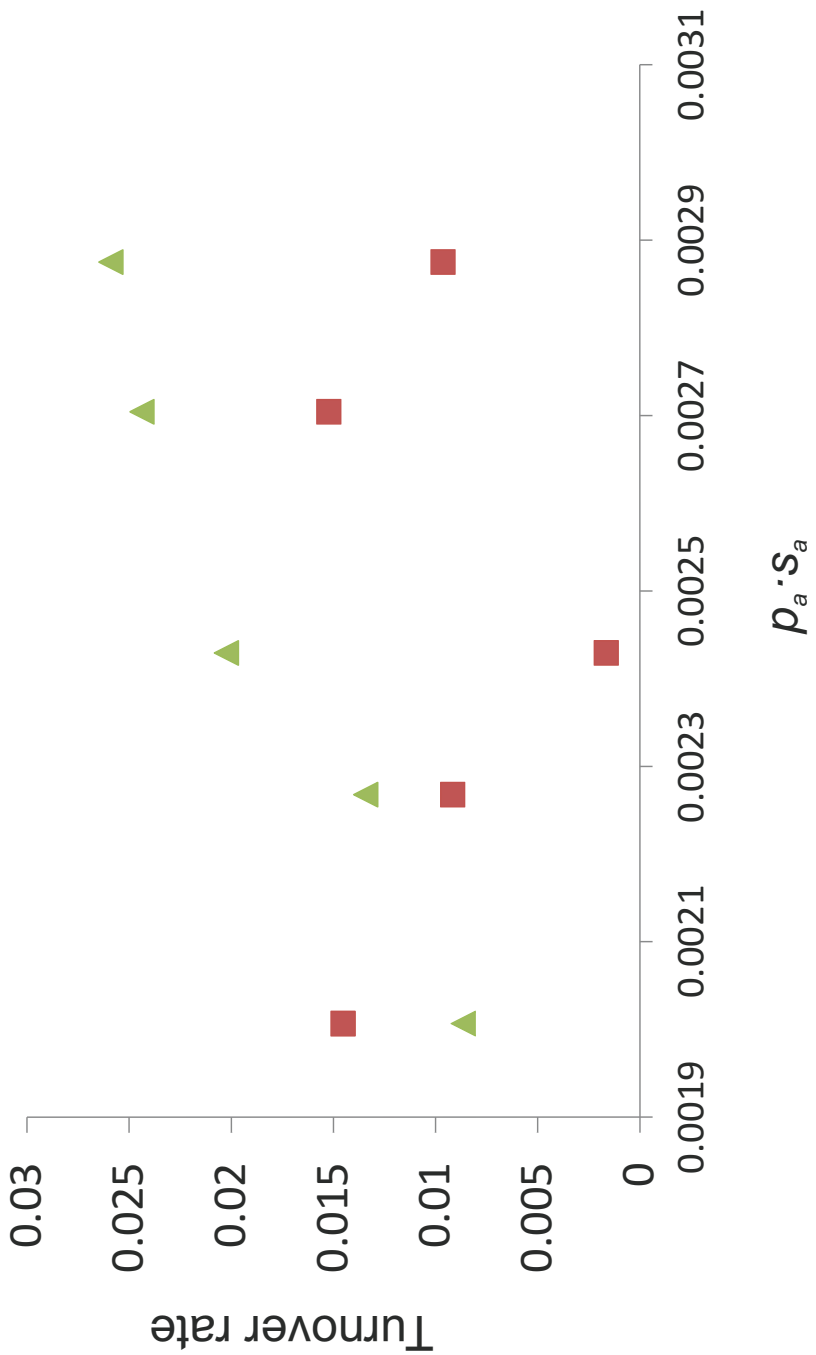
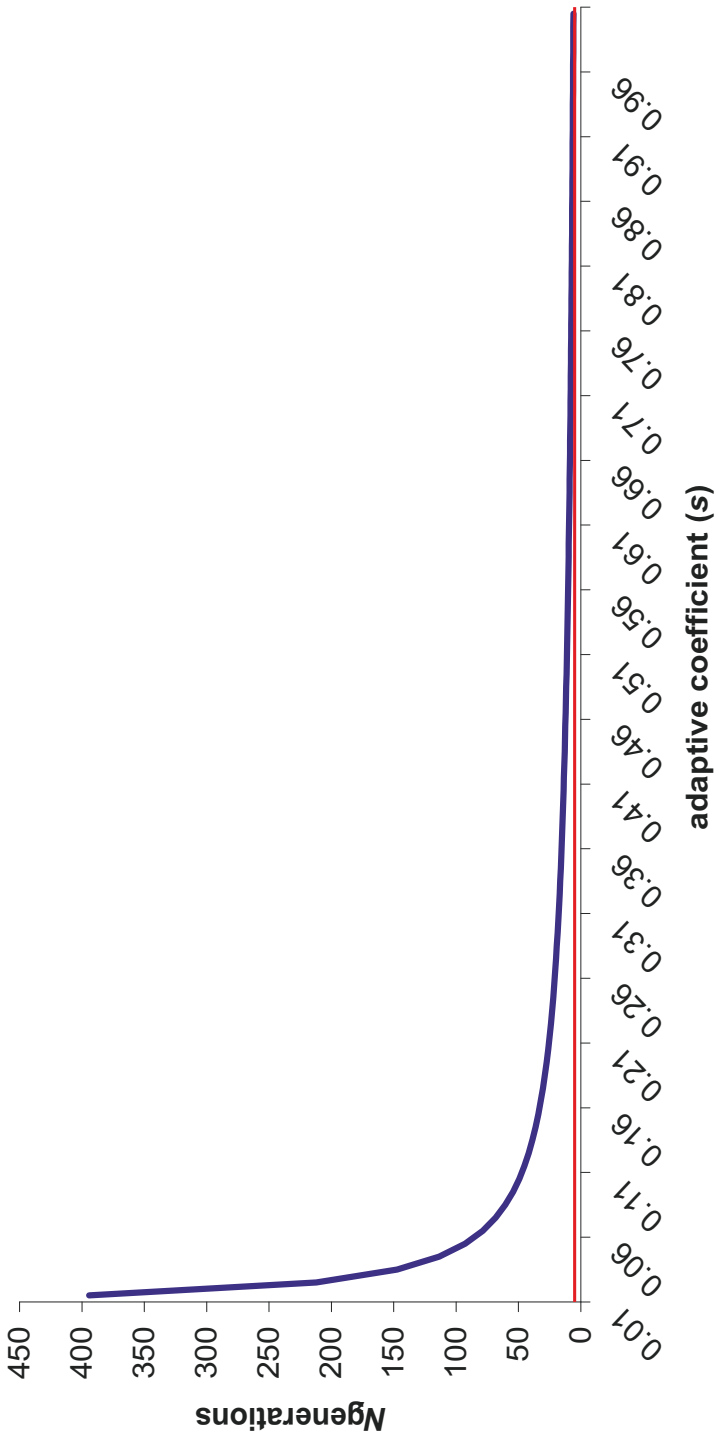

Turnover rate
0.03
0.025
0.02
0.015
0.01
0.005
0
0.0019
0.0021
0.0023
0.0025
0.0027
0.0029
0.0031
$p_a \cdot s_a$

Figure S6


Ngenerations
0
50
100
150
200
250
300
350
400
450
adaptive coefficient (s)
0.01
0.06
0.11
0.16
0.21
0.26
0.31
0.36
0.41
0.46
0.51
0.56
0.61
0.66
0.71
0.76
0.81
0.86
0.91
0.96

4
Discusión

Los primeros estudios de evolución molecular se centraron en la comparación de proteínas por electroforesis, análisis inmunológicos o incluso mediante su propia secuenciación. Uno de los trabajos con mayor repercusión demostró que las diferencias proteicas no eran suficientes para explicar la divergencia anatómica y conductual existente entre humanos y chimpancés [121]. Desde entonces, los cambios en los mecanismos de regulación transcripcional han despertado un creciente interés en la comunidad científica (Figura 4.1). Aun así, su contribución a la divergencia fenotípica sigue siendo una de las cuestiones más controvertidas en biología [19, 21].
La actual disponibilidad de datos moleculares masivos nos dota de una oportunidad sin precedentes para comprender el papel de la selección natural positiva en la evolución de los mecanismos reguladores de la expresión génica. No obstante, la idiosincrasia de estos nuevos datos masivos requiere del previo desarrollo e implementación de potentes herramientas bioinformáticas.

4.1 Implementación de nuevos métodos analíticos

4.1.1 DnaSP v5

Con más de 8000 citas acumuladas entre todas sus versiones, DnaSP [116, 122-125] es uno de los programas más populares en el ámbito de la genética de poblaciones y la evolución molecular. Una de las claves de su éxito radica en la constante actualización de sus funcionalidades. En la versión DnaSP v5 [116], hemos implementado nuevos métodos orientados al estudio de datos masivos de polimorfismo y divergencia nucleotídica, entre los que destaca la capacidad de detectar regiones funcionales por *phylogenetic footprinting* y *phylogenetic shadowing* [126].

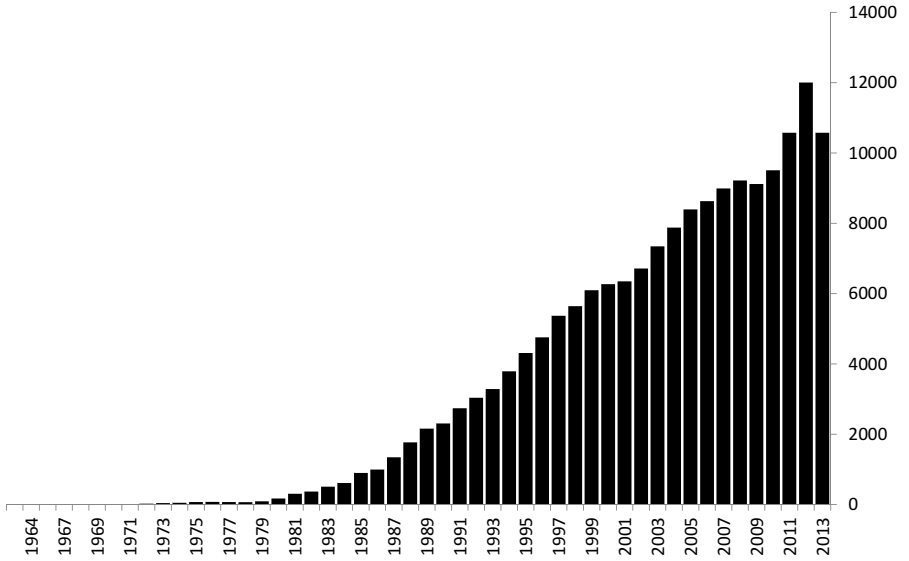


Figura 4.1: Número de publicaciones científicas por año que incluyen la palabra clave '*transcriptional regulation*'.

La idea que subyace al *phylogenetic footprinting* es realmente simple y eficaz: si una región se conserva a lo largo del tiempo, *probablemente* es funcional. Efectivamente, puede haber regiones conservadas no funcionales (falsos positivos). Por ejemplo, si las especies comparadas divergieron recientemente, las regiones genómicas pueden estar conservadas por no haber tenido tiempo de acumular substituciones. Para discernir la constricción funcional de la mera conservación, se desarrollaron las técnicas de *phylogenetic shadowing*. Estás técnicas consideran que una región es funcional *sólo* si está significativamente más conservada de lo esperado (dada la divergencia entre las especies analizadas).

DnaSP implementa una aproximación intermedia. En concreto, detecta aquellas regiones del alineamiento múltiple (MSA) enriquecidas en posiciones conservadas (test de Fisher). La sensibilidad y especificidad del método depende de si el MSA contiene un balance razonable entre regiones funcionales y regiones selectivamente 'neutras'. Si todo el MSA presenta el mismo nivel de conservación (ej. un exón), DnaSP no anotará ningún elemento funcional. Para soslayar esta limitación, el MSA focal se puede concatenar con otro MSA de posiciones 'neutras', lo que reportaría resultados análogos a los obtenidos por *phylogenetic shadowing*.

En esta tesis, hemos utilizado las nuevas implementaciones de DnaSP v5 de forma mayoritariamente prospectiva. Por ejemplo, hemos analizado los niveles y patrones de variabilidad tanto en las regiones *upstream* de los genes

quimiosensoriales, como en sus elementos *cis*-reguladores de la transcripción (CREs).

4.1.2 BadiRate

El análisis comparativo de ganancia y pérdida de genes es fundamental para comprender el papel de la selección natural en moldear el tamaño de las familias multigénicas. Tradicionalmente, la dinámica de las familias multigénicas se ha analizado en un marco de parsimonia, mediante la reconciliación del árbol de genes (AG) con el árbol de especies (AE) [31-33]. Aunque todavía es ampliamente utilizada [98], la reconciliación del AG con el AE es extremadamente sensible a posibles errores metodológicos y violaciones del modelo biológico de nacimiento y muerte de genes (ej. conversión génica o transferencia horizontal) [34, 37, 127]. Hasta la fecha, esto no representaba un problema sustancial, puesto que el análisis estaba circunscrito a unas pocas familias multigénicas que se podían revisar concienzudamente. No obstante, con la creciente disponibilidad de genomas completos, la revisión manual ha dejado de ser una opción viable.

Por ello se han desarrollado nuevos modelos estadísticos que integran parte de la incertidumbre metodológica y biológica [128, 129]. Sin embargo, la parametrización es tan extensa, que muchos no son computacionalmente factibles. Una alternativa más simple es obviar la información de las secuencias génicas y, por tanto, de los sesgos asociados al AG. Intuitivamente se podría pensar que ignorar la secuencia génica es contraproducente, porque se reduciría la relación entre la señal biológica y el ruido estocástico (*signal-to-noise ratio*). No obstante, la variación en el número de copias génicas es -por sí misma- altamente informativa de los procesos evolutivos [130], habiendo demostrado ser de gran utilidad en varios estudios genómicos [131-133].

En el programa BadiRate [134], hemos desarrollado e implementado diferentes modelos estocásticos para estimar la dinámica de ganancia y pérdida de elementos genéticos, ya sean regiones que codifican para proteína o CREs. BadiRate nos ha permitido analizar -por primera vez- la dinámica de los CREs de los genes quimiosensoriales, confiriéndonos una visión evolutiva complementaria al análisis clásico de los niveles y patrones de variación nucleotídica.

Actualmente, estamos trabajando en la incorporación de nuevas funcionalidades, entre las que destacan la capacidad de contrastar un mayor número de hipótesis biológicas, el cálculo de intervalos de confianza, y una interfaz gráfica que facilite su uso por parte de usuarios inexpertos en entornos bioinformáticos complejos.

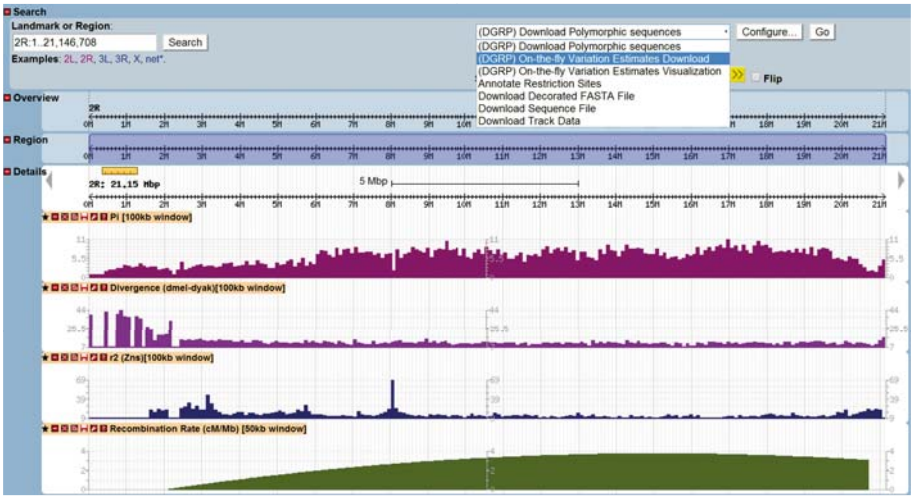


Figura 4.2: Interfaz de popDrowser, mostrando el patrón de diversidad y divergencia nucleotídica (π y K), el desequilibro de ligamiento (r^2), y la tasa de recombinación a lo largo del brazo cromosómico 2R de *D. melanogaster*.

4.1.3 popDrowser

Los primeros esfuerzos de secuenciación genómica se centraron en especies modelo o con algún interés aplicado [110]. No obstante, la reducción de los costes de secuenciación [135] ha potenciado enormemente los estudios de genómica de poblaciones. Por ejemplo, en *D. melanogaster*, existen tres proyectos diferentes: el *Drosophila Population Genomics Project* (DPGP; poblaciones africanas) [136], los 20 genomas europeos [137], y el *Drosophila Genetic Reference Panel* (DGRP; de una única población de Norteamérica; Raleigh, Carolina del Norte) [114].

Como miembros del proyecto DGRP, y de forma consistente con su filosofía (ser un recurso disponible para toda la comunidad científica), hemos desarrollado popDrowser [138]. PopDrowser es una instancia del navegador genómico *Gbrowse* [118] que permite visualizar y analizar el polimorfismo y divergencia (con *D. simulans* y *D. yakuba* como especies *outgroup*) en los genomas de *D. melanogaster* secuenciados por el proyecto DGRP (Figura 4.2). Entre las principales características de popDrowser destaca la posibilidad de analizar estas secuencias genómicas de forma 'remota' (sin necesidad de descargar la información). Además, popDrowser es fácilmente adaptable a cualquier otro proyecto de genómica de poblaciones, lo que le convierte en una herramienta bioinformática muy versátil.

El navegador popDrowser ha sido de gran provecho para el análisis de las regiones *upstream* de los genes quimiosensoriales. Su utilidad radica en el almacenamiento de información ya pre-procesada (alineamientos, diversidad

nucleótidica, etc.), lo que nos ha posibilitado realizar una inspección visual y rápida de la variación nucleotídica en diferentes regiones cromosómicas que incluyen genes quimiosensoriales.

4.2 Evolución de la regulación transcripcional de los genes quimiosensoriales

El sistema quimiosensorial monitoriza las condiciones ecológicas y sociales del ambiente externo, lo que permite modular la conducta del individuo acorde a sus necesidades. En otras palabras, la correcta expresión de los genes quimiosensoriales condiciona la capacidad de discriminar estímulos esenciales (como nutrientes o pareja) y, por tanto, la eficacia biológica de los individuos. Por ello, los mecanismos que regulan la transcripción de los genes quimiosensoriales son -*a priori*- un excelente modelo para estudiar el impacto de la selección natural a nivel molecular.

En esta tesis, hemos estudiado los mecanismos de regulación transcripcional a dos niveles diferentes (pero relacionados). Primero, hemos analizado el impacto de los dominios de la cromatina en la distribución física de los genes que codifican OBPs. Segundo, hemos determinado la contribución relativa de la selección natural en la evolución de las regiones 2Kb *upstream* de los genes del sistema quimiosensorial.

4.2.1 Distribución física de los genes que codifican OBPs

Inferencia de clusters

En el ámbito de la genómica comparada, el término clúster tiene diferentes acepciones, lo que puede motivar cierta confusión. Operacionalmente, un clúster se define como un conjunto de genes parálogos que están físicamente agrupados. Aunque de gran utilidad para comprender la relevancia del entrecruzamiento desigual en la duplicación génica (las copias génicas quedan dispuestas en tándem), esta interpretación carece de sentido funcional *per se*. Al contrario, los clusters de genes (parálogos o no) que se conservan a lo largo del tiempo si pueden tener implicaciones funcionales. Por ejemplo, en *D. melanogaster*, los genes que se transcriben en la cabeza o en el testículo están físicamente agrupados a lo largo del cromosoma. Esta asociación es significativa incluso después de controlar por el efecto del entrecruzamiento desigual en el origen y, por tanto, la distribución de muchos genes parálogos [139].

Con la explosión de la genómica comparada en procariotas, se constató que muchos clusters funcionales pueden experimentar reordenaciones internas

sin que ello comporte consecuencias fenotípicas [140, 141]. Para acomodar estas observaciones, se ha formalizado una nueva definición de clúster, donde la estricta conservación de la colinearidad no es un requisito indispensable, sino que basta con que los genes mantengan su 'vecindad' (el denominado '*gene team model*' [142]).

¿Por qué los genes conservan su 'vecindad' a lo largo del tiempo? La conservación de los clusters resulta -principalmente- de la interacción de dos fuerzas evolutivas: (i) la mutación, ya que las regiones poco propensas a sufrir reordenaciones cromosómicas mantendrán su colinearidad a lo largo del tiempo [96, 97, 143, 144], y (ii) la selección natural, que puede mantener los *clusters* de forma activa por diversos motivos, como facilitar la co-regulación transcripcional de sus integrantes. Para discernir entre ambas hipótesis, hemos aplicado una nueva y eficiente implementación del '*gene team model*' cuya filosofía es similar a la del *phylogenetic shadowing* [145]: se considera que los clusters están sometidos a constricción funcional *sólo* si están significativamente más conservados de lo esperado (dada la tasa de reordenamiento cromosómico en *Drosophila*).

La sensibilidad y especificidad para inferir clusters bajo constricción funcional incrementa con la divergencia entre las especies comparadas, así como también con la calidad de sus ensamblajes y anotaciones genómicas. Intuitivamente, se podría pensar que nuestro modelo de estudio (las 12 especies de *Drosophila*) atesora un excelente balance (representatividad filogenética vs. calidad de la información genómica) para realizar análisis de este tipo. Sin embargo, hemos apreciado que la divergencia entre las 12 *Drosophila* no proporciona suficiente potencia estadística para detectar los clusters más pequeños (de dos o tres miembros) que estén activamente mantenidos por la selección natural. Esto tiene importantes connotaciones, ya que se ha postulado que existen dos tipos de clusters, los denominados 'pequeños' (constituidos por pocos miembros que están filogenéticamente relacionados y altamente co-regulados), y los denominados 'grandes' (formados por genes que se expresan en múltiples condiciones, o genes *housekeeping*) [146]. Aunque pueden estar parcialmente sesgados, nuestros resultados no pretenden minimizar la relevancia de los clusters 'pequeños', sino poner de manifiesto la importancia de los 'grandes' en la distribución física de los genes que codifican OBPs.

Clusters que incluyen OBPs

Hemos identificado un total de 31 clusters que -en promedio- están conservados en 5.9 especies y contienen 8.3 genes, 1.7 de los cuales codifica OBPs. Más allá de la media, el repertorio génico por clúster es muy variable. Por una parte, el clúster con mayor densidad génica abarca la región 20284679-20292009 del cromosoma X de *D. melanogaster*, e incluye cinco genes, cuatro de los cuales codifican OBPs (*Obp19a-d*). Por otra parte, uno de los clusters bajo mayor

constricción funcional contiene 20 genes, pero tan sólo uno de ellos codifica una OBP (*lush*). Los 19 genes restantes están involucrados en procesos aparentemente relacionados, como *Shal* (un canal de potasio), *asf1* (morfogénesis de dendritas), y *tey* (transmisión sináptica).
¿Qué función desempeñan los genes que están *clusterizados* con las OBPs? Hemos encontrado que -aparte de las propias funciones olfativas- estos clusters incluyen genes que típicamente se transcriben en el sistema nervioso, como canales de sodio, receptores de neurotransmisores y, especialmente, proteínas integrales de la membrana celular (test de Fisher; $P = 1.06e^{-10}$).

Causas de la conservación de los clusters

La co-regulación transcripcional permite orquestar la expresión génica frente a estímulos externos, y en diferentes tejidos o estadios del desarrollo. El caso más ilustrativo es -probablemente- el del operón bacteriano *lac*, que coordina la transcripción policistrónica de tres enzimas, dos de los cuales están involucrados en la degradación de la lactosa [147]. Salvando ciertas diferencias moleculares, los clusters eucariotas también pueden facilitar la co-regulación transcripcional de sus miembros. En vertebrados, por ejemplo, los genes que codifican los factores de transcripción Hox están físicamente agrupados, y además en el mismo orden en el que se transcriben durante el desarrollo [148-150].
Nuestros análisis multivariantes han revelado que la constricción funcional de los 31 clusters focales correlaciona positivamente con la amplitud (EB) y el ruido (EN) transcripcional de sus integrantes (*Path analysis*; $P = 0.004$ y $P = 0.043$, respectivamente). Este resultado es -aparentemente- paradójico, ya que el EN *suele* ser deletéreo [102, 103]. Sin embargo, en algunas circunstancias, el EN puede generar una plasticidad fenotípica beneficiosa, especialmente si afecta la expresión de proteínas que están en contacto directo con ambientes externos cambiantes, como transportadores extracelulares (ej. OBPs) o proteínas integrales de membrana (ej. canales de sodio) [151, 152]. Así, las fluctuaciones estocásticas en la abundancia de OBPs podrían inducir respuestas conductuales dispares frente al mismo estímulo [153, 154], una variación fenotípica que quizás juegue un papel fundamental en la percepción olfativa.

Mecanismos de co-regulación transcripcional

Durante las primeras décadas del siglo XX, el estudio de los cromosomas politénicos de *Drosophila* reveló la existencia de dos estados bien diferenciados de la cromatina: la eucromatina y la heterocromatina. Aunque ya se habían constatado diferencias en cuanto a la densidad génica, no fue hasta 1930 cuando Hermann Muller demostró que el estado de la cromatina también condiciona la transcripción génica (el denominado efecto de posición) [155]. Actualmente,

las nuevas técnicas experimentales han permitido identificar un amplio abanico de estados de la cromatina, cada uno con sus pecularidades transcripcionales [112].

Hemos demostrado que el nivel de constricción funcional de los 31 clusters focales está asociado con un estado de la cromatina denominado 'elongación de la transcripción' (correlación de Spearman; $P = 0.006$) y, de forma muy interesante, con la unión de la proteína quinasa JIL-1 (correlación de Spearman; $P < 0.001$). La JIL-1 interactúa físicamente con Lam_{D0} [106] y Chromator [108], dos proteínas estructuralmente asociadas con la membrana nuclear que ya han sido -directa o indirectamente- relacionadas con el mantenimiento de los *clusters* [109, 156]. Entre otras funciones, la JIL-1 libera la ARN polimerasa pausada en la región promotora [157], lo que produce una ráfaga de elongación de la transcripción génica que incrementa el EN [158].

La arquitectura de la región promotora puede tener un papel fundamental en pausar la actividad de la ARN polimerasa. En este contexo cabe destacar que se han identificado dos tipos de promotores, los denominados *peaked* (normalmente inician la transcripción a partir del mismo nucleótido) y los denominados *broad* (desde posiciones más dispersas) [159]. De forma consistente con los anteriores resultados, también hemos encontrado una asociación significativa entre la proporción de promotores del tipo *broad* y la constricción funcional del clúster (correlación de Spearman; $P = 0.044$). Dado que la arquitectura del tipo *broad* está asociada con un elevado EB, con un posicionamiento particular de los nucleosomas y con una composición distintiva de CREs [105, 159], también hemos estudiado la contribución de la selección natural en la evolución de las regiones *upstream* de las principales familias multigénicas del sistema quimiosensorial.

4.2.2 Evolución de las regiones *upstream*

Historia demográfica de las poblaciones norteamericanas de *D. melanogaster*

Al igual que otras especies comensales de humanos, la distribución geográfica de *D. melanogaster* está íntimamente ligada a nuestra historia demográfica. Así, en América del Norte, el primer espécimen de *D. melanogaster* se capturó en 1875 (Nueva York), probablemente trasladada de forma involuntaria por el comercio naval con Europa y/o Centroamérica [160]. Sólo 30 años después, la mosca del vinagre ya estaba catalogada como el díptero más común de todo el continente [161] (Figura 4.3). Como la colonización del hábitat norteamericano es tan reciente, la huella molecular de la selección adaptativa (donde el sistema quimiosensorial puede tener un papel fundamental) podría ser fácilmente detectable en el genoma.

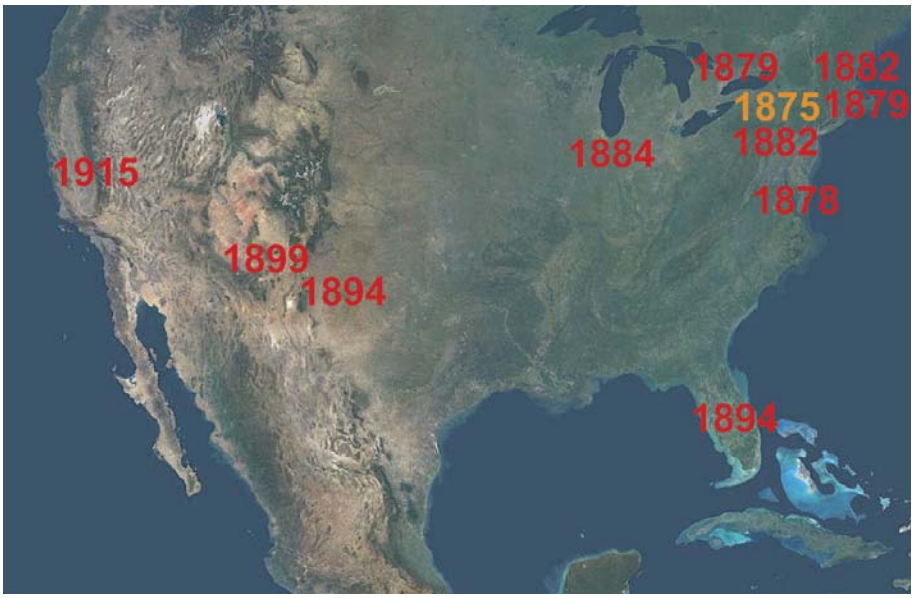


Figura 4.3: Colonización de América del Norte por *D. melanogaster*. Las fechas indican el año en el que la especie se capturó por primera vez en cada población. Fuente: [161].

Los eventos demográficos -sin embargo- moldean la frecuencia de las mutaciones que segregan en la población, pudiendo dejar una huella molecular similar a la de la selección natural [162]. La aproximación más popular para discernir el impacto de la selección natural es la extensión del test de McDonald-Kreitman (eMK) [6, 163]. A pesar de su popularidad, esta aproximación no está exenta de problemas. La eMK asume que la selección natural negativa es suficientemente eficaz como para purgar las mutaciones ligeramente deletéreas (MLDs) de la población, una suposición que podría no cumplirse en las poblaciones norteamericanas de *D. melanogaster*. Primero, el efecto fundador asociado a cualquier proceso de colonización implica un bajo N_e y, por tanto, una reducción de la eficacia de la selección natural (algunas MLDs pueden estar segregando a baja frecuencia) [164, 165]. Segundo, después del cuello de botella inicial, la población de *D. melanogaster* se expandió rápidamente por Norteamérica (Figura 4.3), lo que pudo llevar a la fijación de algunas MLDs.

En esta tesis, hemos utilizado un potente y novedoso método de genética de poblaciones [8, 166] que primero controla por el efecto de la historia demográfica, para luego estimar los principales parámetros selectivos: el efecto de las mutaciones deletéreas, la tasa de substitución adaptativa (α [167]), y α normalizada por la tasa de substitución neutra (ω_α) [168].

Impacto de la selección *darwiniana* en las regiones *upstream*

Hemos estimado que el 95 % de las substituciones ocurridas en las 2Kb *upstream* de los genes quimiosensoriales se han fijado por el efecto de la selección natural positiva ($\alpha \approx 0.95$), una estima muy superior al 50 % inferido en las regiones no codificadoras de *D. simulans* [169]. ¿Cómo se concilian estos resultados tan dispares? La respuesta viene dada -otra vez- por la historia demográfica de la población. Generalmente, el efecto fundador (unos pocos individuos colonizan un nuevo hábitat) comporta una reducción de la variabilidad genética, que sólo se recupera con el paso del tiempo. En esta población, sin embargo, la recuperación del cuello de botella es tan reciente que apenas las mutaciones ventajosas han tenido tiempo para fijarse.
Si α depende de cada historia demográfica, ¿cómo podemos comparar la contribución de la selección natural en la evolución de diferentes poblaciones? La poca interpretabilidad de α (en términos selectivos) se puede solventar normalizándola por la tasa de substitución neutra (una normalización denominada ω_α) [168]. En este sentido, nuestras estimas de ω_α (desde $\omega_\alpha = 0.266$ en aIRs hasta $\omega_\alpha = 0.440$ en ORs) son similares a las obtenidas para 373 regiones de *D. melanogaster* que codifican proteína ($\omega_\alpha \approx 0.4$) [168], lo que soporta la idea de que la estima de $\alpha \approx 0.95$ es un mero subproducto de la historia demográfica.
La anterior comparativa de ω_α contradice, además, una de nuestras hipótesis de partida: las regiones *upstream* de los genes quimiosensoriales no son una de las principales dianas de la selección *darwiniana* (estimada como ω_α). De hecho, ni la proporción (p_a) ni el coeficiente de la selección adaptativa (s_a) son mayores en las regiones *upstream* de los genes quimiosensoriales que en el resto de regiones *upstream* del genoma (análisis de muestreo con reemplazamiento, o *boostrap*; $P > 0.05$, para todas las familias quimiosensoriales analizadas).

Impacto de la selección *darwiniana* en los CREs

El análisis indiscriminado de 2Kb (región *upstream*) puede inducir interpretaciones erróneas acerca del impacto de la selección natural en la evolución transcripcional, puesto que estas regiones incluyen muchos nucleótidos que no participan en el control de la expresión génica. Ciertamente, el número de CREs necesarios para conferir un correcto patrón transcripcional es muy heterogéneo, siendo aparentemente menor en los genes quimiosensoriales que en el resto (test de Wilcoxon; $P < 0.05$ para todas las familias analizadas, excepto para los aIRs).
Para examinar el impacto de la selección natural en las regiones que si controlan la transcripción génica, hemos comparado el polimorfismo y la divergencia (π/K) de los CREs con el resto de regiones *upstream*. La relación π/K es menor en los CREs que en las demás posiciones (test de Wilcoxon; $P <$

$2.2e^{-16}$), principalmente porque la selección natural negativa purga las mutaciones deletéreas, produciendo un déficit relativo de π (test de Wilcoxon; $P < 2.2e^{-16}$).
Aunque la selección purificadora juega un papel fundamental en su conservación [170], el contenido de CREs es muy variable entre genes ortólogos [171-173]. Tradicionalmente, esta elevada dinámica se ha interpretado como resultado de la deriva genética y de las mutaciones compensatorias [174, 175]. No obstante, recientemente se han encontrado algunas evidencias discrepantes, que sugieren una mayor contribución de la selección natural positiva [176]. Gracias al desarrollo e implementación de nuevos modelos de ganancia y pérdida de CREs en BadiRate [134], hemos podido contrastar estas dos hipótesis en un marco probabilístico. De forma notable, hemos detectado una leve (pero significativa) asociación entre el impacto de la selección adaptativa y los eventos de ganancia (test de Wilcoxon; $P = 0.0472$) y pérdida (test de Wilcoxon; $P = 0.0316$) de CREs de los genes quimiosensoriales, lo cual es consistente con un impacto de la selección *darwiniana* en la evolución de su regulación transcripcional.

Evolución de las familias multigénicas del sistema quimiosensorial de *Drosophila*

La creciente disponibilidad de genomas completos está incrementando sustancialmente nuestro conocimiento de la evolución de las familias multigénicas del sistema quimiosensorial [80], lo cual es absolutamente imprescindible para comprender su origen y función. En este sentido, la comparación del impacto de la selección natural revela dos modos de evolución bien diferenciados (Figura 4.4). Por una parte, los aIRs y las CSPs están sometidos a una fuerte constricción, lo cual es consistente con nuestro conocimiento evolutivo y funcional de ambas familias: tienen un origen antiguo y juegan un papel basal en la quimiopercepción, así como también en otros procesos biológicos [61, 67, 75].
Las OBPs, los OR y los GRs evolucionan -por otra parte- más rápidamente, y de forma parecida entre sí. Un aspecto interesante atañe al impacto de la selección natural positiva en las regiones *upstream*; a diferencia de lo observado para λ y ω (donde las OBPs y los ORs parecen estar sujetas a presiones selectivas parecidas[80]; Figura 4.4) , las estimas de ω_α son significativamente más elevadas en los ORs (intervalos de confianza estimados por *bootstrap*; $\omega_\alpha = 0.442$ [0.399-0.471]) y GRs ($\omega_\alpha = 0.415$ [0.386-0.445]) que en las OBPs ($\omega_\alpha = 0.349$ [0.311-0.383]).
Aunque aún faltan evidencias al respecto, esta discrepancia evolutiva podría estar vinculada con el impacto diferencial de la cromatina en la regulación transcripcional de las OBPs y de los ORs. Efectivamente, hemos encontrado

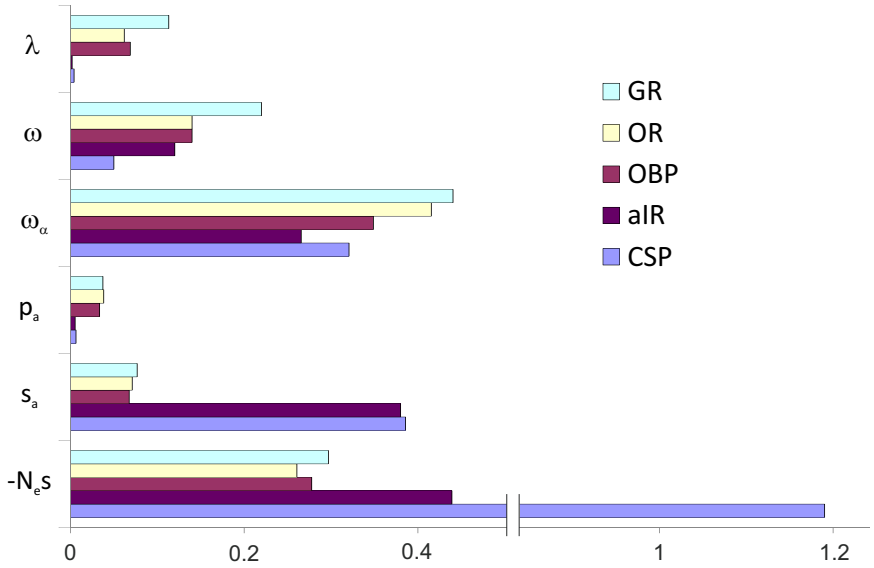


Figura 4.4: Comparación del impacto de la selección natural en la evolución de las principales familias del sistema quimiosensorial. Dinámica de ganancia y pérdida de genes (λ); constricción funcional en las regiones codificadoras (ω); presión selectiva en las regiones *upstream* (ω_α); proporción (p_a) y coeficiente de selección (s_a) de las mutaciones beneficiosas; efecto de las mutaciones deletéreas ($-N_e s$). Por propósitos ilustrativos, λ ha sido multiplicada por 10, y $-N_e s$ divido entre 100. Los datos de λ y ω han sido extraídos de [80].

que los clusters que incluyen OBPs están localizados en regiones cromosómicas caracterizadas por un estado de la cromatina denominado 'elongación de la transcripción', lo que a su vez repercute en la amplitud y el ruido transcripcional de sus integrantes. De cualquier modo, no cabe duda que la selección natural (negativa y positiva) ha jugado un papel fundamental en la evolución de los mecanismos que regulan la transcripción de los genes quimiosensoriales, tanto los que implican elementos *cis*-reguladores (CREs), como los dominios de la cromatina.

5
Conclusiones

1. Hemos implementado nuevos métodos de genética de poblaciones y evolución molecular en el programa DnaSP, entre los que destaca la inferencia de regiones funcionales por *phylogenetic footprinting* y *phylogenetic shadowing*, y el análisis automatizado de los niveles y patrones de variabilidad en múltiples *loci*.

2. Hemos implementado popDrowser, una instancia del navegador genómico *Gbrowse* especialmente diseñada para proyectos de genómica de poblaciones, como el *Drosophila Genetic Reference Panel* (DGRP). La versatilidad de popDrowser radica en la posibilidad de realizar análisis de forma 'remota', lo que facilita la visualización inmediata de los resultados con el resto de anotaciones integradas en el propio navegador.

3. Hemos desarrollado nuevos modelos estocásticos para analizar la ganancia y pérdida de elementos genéticos, ya sean genes eucariotas (que mayoritariamente se originan por duplicación de ADN), procariotas (donde la transferencia horizontal puede jugar un papel importante) o incluso elementos *cis*-reguladores de la transcripción (que pueden tener un origen *de novo*).

4. Hemos implementado diversos modelos estocásticos para el análisis de la ganancia y pérdida de elementos genéticos en el programa BadiRate. Este programa proporciona el marco filogenético y probabilístico apropiado para contrastar diferentes hipótesis evolutivas, como expansiones y contracciones de familias multigénicas en linajes concretos.

5. Los genes que codifican OBPs están activamente mantenidos en clusters por la acción de la selección natural. Estos clusters incluyen además otros genes, entre los que destacan los que codifican proteínas de la

membrana celular.

6. Los análisis multivariantes revelan que la conservación de estos clusters está relacionada con: (i) la amplitud y el ruido transcripcional de sus integrantes; (ii) el estado de la cromatina (*'transcription elongation'* y con la unión de la proteína JIL-1); (iii) la arquitectura de la región promotora de sus genes.
7. Las inferencias mediante potentes métodos de genética de poblaciones indican que la selección natural (negativa y positiva) contribuye significativamente a la evolución de las regiones *upstream* de los genes del sistema quimiosensorial.
8. Las regiones *upstream* de los aIRs y las CSPs están sometidas a una importante constricción funcional, mientras que las de los ORs y los GRs han experimentado un mayor impacto de la selección *darwiniana*.
9. La huella molecular de la selección natural no es homogénea a lo largo de la región *upstream*, sino que mayoritariamente solapa con los CREs.
10. El impacto de la selección natural positiva correlaciona positivamente con la dinámica de ganancia y pérdida de los CREs que regulan la transcripción de los genes quimiosensoriales.

Bibliografía

Bibliografía

[1] Gordon Campbell. "Empedocles". In: *Internet Encyclopedia of Philosophy* (2005).

[2] Dirk L. Couprie. "Anaximander". In: *Internet Encyclopedia of Philosophy* (2005).

[3] Charles. Bonnet. *Considerations sur les corps organises*. Chez Marc-Michel Rey, 1762.

[4] C. Darwin. *On the Origin of Species by Means of Natural Selection, Or, The Preservation of Favoured Races in the Struggle for Life*. J. Murray, 1859.

[5] Svante Paabo. "The mosaic that is our genome". In: *Nature* 421.6921 (Jan. 2003), pp. 409–412.

[6] J. H. McDonald and M. Kreitman. "Adaptive protein evolution at the Adh locus in Drosophila". In: *Nature* 351.6328 (1991).

[7] Richard R. Hudson, Martin Kreitman, and Montserrat Aguade. "A Test of Neutral Molecular Evolution Based on Nucleotide Data". In: *Genetics* 116.1 (May 1987).

[8] Adrian Schneider, Brian Charlesworth, Adam Eyre-Walker, and Peter D Keightley. "A method for inferring the rate of occurrence and fitness effects of advantageous mutations". In: *Genetics* 189.4 (Dec. 2011).

[9] Z. Yang, R. Nielsen, N. Goldman, and A. M. Pedersen. "Codon-substitution models for heterogeneous selection pressure at amino acid sites". In: *Genetics* 155 (2000).

[10] Z. Yang and W. J. Swanson. "Codon-substitution models to detect adaptive evolution that account for heterogeneous selective pressures among site classes". In: *Mol Biol Evol* 19 (2002).

[11] Consortium Drosophila 12 Genomes, A. G. Clark, M. B. Eisen, D. R. Smith, C. M. Bergman, et al. "Evolution of genes and genomes on the Drosophila phylogeny". In: *Nature* 450 (2007).

[12] Carlos D. Bustamante, Adi Fledel-Alon, Scott Williamson, Rasmus Nielsen, Melissa Todd Hubisz, et al. "Natural selection on protein-coding genes in the human genome". In: *Nature* 437.7062 (Oct. 2005).

[13] Daniel L. Halligan, Athanasios Kousathanas, Rob W. Ness, Bettina Harr, Lel Eory, et al. "Contributions of Protein-Coding and Regulatory Change to Adaptive Molecular Evolution in Murid Rodents". In: *PLoS Genet* 9.12 (Dec. 2013).

[14] E S Lander, L M Linton, B Birren, C Nusbaum, M C Zody, et al. "Initial sequencing and analysis of the human genome". In: *Nature* 409.6822 (Feb. 2001).

[15] International Human Genome Sequencing Consortium. "Finishing the euchromatic sequence of the human genome". In: *Nature* 431.7011 (Oct. 2004).

[16] Erica Sodergren, George M. Weinstock, Eric H. Davidson, R. Andrew Cameron, Richard A. Gibbs, et al. "The Genome of the Sea Urchin Strongylocentrotus purpuratus". In: *Science* 314.5801 (Nov. 2006).

[17] Stephen A Goff, Darrell Ricke, Tien-Hung Lan, Gernot Presting, Ronglin Wang, et al. "A draft sequence of the rice genome (Oryza sativa L. ssp. japonica)". In: *Science* 296.5565 (Apr. 2002).

[18] Patricia J. Wittkopp and Gizem Kalay. "Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence". In: *Nat Rev Genet* 13.1 (Jan. 2012).

[19] Sean B Carroll. "Evolution at two levels: on genes and form". In: *PLoS Biol.* 3.7 (July 2005).

[20] Dallas M Swallow. "Genetics of lactase persistence and lactose intolerance". In: *Annu. Rev. Genet.* 37 (2003).

[21] Hopi E Hoekstra and Jerry A Coyne. "The locus of evolution: evo devo and the genetics of adaptation". In: *Evolution* 61.5 (May 2007).

[22] K Luger, A W Mäder, R K Richmond, D F Sargent, and T J Richmond. "Crystal structure of the nucleosome core particle at 2.8 A resolution". In: *Nature* 389.6648 (Sept. 1997).

[23] G. Li and D. Reinberg. "Chromatin higher-order structures and gene regulation". In: *Curr Opin Genet Dev* 21.2 (2011).

[24] D J Steger and J L Workman. "Remodeling chromatin structures for transcription: what happens to the histones?" In: *Bioessays* 18.11 (Nov. 1996).

[25] S. Thomas, X. Y. Li, P. J. Sabo, R. Sandstrom, R. E. Thurman, et al. "Dynamic reprogramming of chromatin accessibility during Drosophila embryo development". In: *Genome Biol* 12.5 (2011).

[26] Marissa Vignali, Ahmed H. Hassan, Kristen E. Neely, and Jerry L. Workman. "ATP-Dependent Chromatin-Remodeling Complexes". In: *Mol. Cell. Biol.* 20.6 (Mar. 2000).

[27] Roger A Hoskins, Christopher D Smith, Joseph W Carlson, A Bernardo Carvalho, Aaron Halpern, et al. "Heterochromatic sequences in a Drosophila whole-genome shotgun assembly". In: *Genome Biol.* 3.12 (2002).

[28] M. Nei and A. P. Rooney. "Concerted and birth-and-death evolution of multigene families". In: *Annu Rev Genet* 39 (2005).

[29] Bernard Conrad and Stylianos E Antonarakis. "Gene duplication: a drive for phenotypic diversity and cause of human disease". In: *Annu Rev Genomics Hum Genet* 8 (2007).

[30] H. Innan and F. Kondrashov. "The evolution of gene duplications: classifying and distinguishing between models". In: *Nat Rev Genet* 11.2 (2010).

[31] Morris Goodman, John Czelusniak, G. William Moore, A. E. Romero-Herrera, and Genji Matsuda. "Fitting the Gene Lineage into its Species Lineage, a Parsimony Strategy Illustrated by Cladograms Constructed from Globin Sequences". In: *Systematic Zoology* 28.2 (June 1979).

[32] Benjamin Vernot, Maureen Stolzer, Aiton Goldman, and Dannie Durand. "Reconciliation with non-binary species trees". In: *J. Comput. Biol.* 15.8 (Oct. 2008).

[33] Bret R. Larget, Satish K. Kotha, Colin N. Dewey, and Cécile Ané. "BUCKy: Gene tree/species tree reconciliation with Bayesian concordance analysis". In: *Bioinformatics* 26.22 (Nov. 2010).

[34] Wayne P. Maddison. "Gene Trees in Species Trees". In: *Syst Biol* 46.3 (Sept. 1997).

[35] Claudio Casola, Carrie L. Ganote, and Matthew W. Hahn. "Nonallelic Gene Conversion in the Genus Drosophila". In: *Genetics* 185.1 (May 2010).

[36] Krister M. Swenson and Nadia El-Mabrouk. "Gene trees and species trees: irreconcilable differences". In: *BMC Bioinformatics* 13.19 (Dec. 2012).

[37] M. W. Hahn. "Bias in phylogenetic tree reconciliation methods: implications for vertebrate genome evolution". In: *Genome Biol* 8.7 (2007).

[38] M. W. Hahn, T. De Bie, J. E. Stajich, C. Nguyen, and N. Cristianini. "Estimating the tempo and mode of gene family evolution from comparative genomic data". In: *Genome Res* 15.8 (2005).

[39] M. Csuros and I. Miklos. "Streamlining and large ancestral genomes in Archaea inferred with a phylogenetic birth-and-death model". In: *Mol Biol Evol* 26.9 (2009).

[40] J.W. Bradbury and S.L. Vehrencamp. *Principles of Animal Communication*. Sinauer Associates, Incorporated, 2011.

[41] M. J. Krieger and K. G. Ross. "Identification of a major gene regulating complex social behavior". In: *Science* 295 (2002).

[42] T. Matsuo, S. Sugaya, J. Yasukawa, T. Aigaki, and Y. Fuyama. "Odorant-binding proteins OBP57d and OBP57e affect taste perception and host-plant preference in Drosophila sechellia". In: *PLoS Biol* 5.5 (2007).

[43] C Smadja and R K Butlin. "On the scent of speciation: the chemosensory system and its role in premating isolation". In: *Heredity (Edinb)* 102.1 (Jan. 2009).

[44] Kenta Asahina, Viktoryia Pavlenkovich, and Leslie B. Vosshall. "The survival advantage of olfaction in a competitive environment". In: *Curr Biol* 18.15 (Aug. 2008).

[45] Marcus C Stensmyr, Susanne Erland, Eric Hallberg, Rita Wallén, Peter Greenaway, et al. "Insect-like olfactory adaptations in the terrestrial giant robber crab". In: *Curr. Biol.* 15.2 (Jan. 2005).

[46] Yoshihito Niimura. "On the Origin and Evolution of Vertebrate Olfactory Receptor Genes: Comparative Genome Analysis Among 23 Chordate Species". In: *Genome Biol Evol* 1 (Jan. 2009).

[47] Anna-Sara Krång, Markus Knaden, Kathrin Steck, and Bill S Hansson. "Transition from sea to land: olfactory function and constraints in the terrestrial hermit crab Coenobita clypeatus". In: *Proc. Biol. Sci.* 279.1742 (Sept. 2012).

[48] T S McClintock and B W Ache. "Ionic currents and ion channels of lobster olfactory receptor neurons." In: *The Journal of General Physiology* 94.6 (1989), pp. 1085–1099.

[49] P. Pelosi. "Perireceptor events in olfaction". In: *J Neurobiol* 30 (1996).

[50] J G Hildebrand and G M Shepherd. "Mechanisms of olfactory discrimination: converging evidence for common principles across phyla". In: *Annu. Rev. Neurosci.* 20 (1997).

[51] Monika Stengl and Nico W Funk. "The role of the coreceptor Orco in insect olfactory transduction". In: *J. Comp. Physiol. A Neuroethol. Sens. Neural. Behav. Physiol.* 199.11 (Nov. 2013).

[52] Kensaku Mori and Hitoshi Sakano. "How Is the Olfactory Map Formed and Interpreted in the Mammalian Brain?" In: *Annual Review of Neuroscience* 34.1 (2011).

[53] Leslie B Vosshall and Reinhard F Stocker. "Molecular architecture of smell and taste in Drosophila". In: *Annu. Rev. Neurosci.* 30 (2007).

[54] P J Clyne, C G Warr, M R Freeman, D Lessing, J Kim, et al. "A novel family of divergent seven-transmembrane proteins: candidate odorant receptors in Drosophila". In: *Neuron* 22.2 (Feb. 1999).

[55] Q Gao and A Chess. "Identification of candidate Drosophila olfactory receptors from genomic DNA sequence". In: *Genomics* 60.1 (Aug. 1999).

[56] D. S. Hekmat-Scafe, C. R. Scafe, A. J. McKinney, and M. A. Tanouye. "Genome-wide analysis of the odorant-binding protein gene family in Drosophila melanogaster". In: *Genome Res* 12.9 (2002).

[57] Sergio Angeli, Francesca Ceron, Andrea Scaloni, Maria Monti, Gaia Monteforti, et al. "Purification, structural characterization, cloning and immunocytochemical localization of chemoreception proteins from Schistocerca gregaria". In: *European Journal of Biochemistry* 262.3 (1999).

[58] R. R. H. Anholt and T. I. Williams. "The Soluble Proteome of the Drosophila Antenna". In: *Chemical Senses* 35.1 (Nov. 2009).

[59] A. Xu, S.-K. Park, S. D'Mello, E. Kim, Q. Wang, et al. "Novel genes expressed in subsets of chemosensory sensilla on the front legs of male Drosophila melanogaster". In: *Cell Tissue Res* 307.3 (Mar. 2002).

[60] Su K. Park, Kevin J. Mann, Heping Lin, Elena Starostina, Aaron Kolski-Andreaco, et al. "A Drosophila Protein Specific to Pheromone-Sensing Gustatory Hairs Delays Males' Copulation Attempts". In: *Current Biology* 16.11 (June 2006).

[61] F. G. Vieira and J. Rozas. "Comparative Genomics of the Odorant-Binding and Chemosensory Protein Gene Families across the Arthropoda: Origin and evolutionary history of the chemosensory system". In: *Genome Biol Evol* 3 (2011).

[62] Malini Manoharan, Matthieu Ng Fuk Chong, Aurore Vaïtinadapoulé, Etienne Frumence, Ramanathan Sowdhamini, et al. "Comparative Genomics of Odorant Binding Proteins in Anopheles gambiae, Aedes aegypti, and Culex quinquefasciatus". In: *Genome Biol Evol* 5.1 (Jan. 2013).

[63] P. Xu, R. Atkinson, D. N. Jones, and D. P. Smith. "Drosophila OBP LUSH is required for activity of pheromone-sensitive neurons". In: *Neuron* 45.2 (2005).

[64] Carolina Gomez-Diaz, Jaime H. Reina, Christian Cambillau, and Richard Benton. "Ligands for Pheromone-Sensing Neurons Are Not Conformationally Activated Odorant Binding Proteins". In: *PLoS Biol* 11.4 (Apr. 2013).

[65] V. R. Chintapalli, J. Wang, and J. A. Dow. "Using FlyAtlas to identify better Drosophila melanogaster models of human disease". In: *Nat Genet* 39.6 (2007).

[66] J. Pelletier and W. S. Leal. "Genome analysis and expression patterns of odorant-binding proteins from the Southern House mosquito Culex pipiens quinquefasciatus". In: *PLoS One* 4.7 (2009).

[67] A Nomura, K Kawasaki, T Kubo, and S Natori. "Purification and localization of p10, a novel protein that increases in nymphal regenerating legs of Periplaneta americana (American cockroach)". In: *Int. J. Dev. Biol.* 36.3 (Sept. 1992).

[68] Yehuda Ben-Shahar, Beika Lu, Daniel M. Collier, Peter M. Snyder, Mikael Schnizler, et al. "The Drosophila Gene CheB42a Is a Novel Modifier of Deg/ENaC Channel Function". In: *PLoS ONE* 5.2 (Feb. 2010).

[69] H. M. Robertson, C. G. Warr, and J. R. Carlson. "Molecular evolution of the insect chemoreceptor gene superfamily in Drosophila melanogaster". In: *Proc Natl Acad Sci U S A* 100 Suppl 2 (2003).

[70] R. Benton, K. S. Vannice, C. Gomez-Diaz, and L. B. Vosshall. "Variant ionotropic glutamate receptors as chemosensory receptors in Drosophila". In: *Cell* 136.1 (2009).

[71] Richard Benton, Silke Sachse, Stephen W Michnick, and Leslie B Vosshall. "Atypical Membrane Topology and Heteromeric Function of Drosophila Odorant Receptors In Vivo". In: *PLoS Biol* 4.2 (Jan. 2006).

[72] Hui-Jie Zhang, Alisha R. Anderson, Stephen C. Trowell, A-Rong Luo, Zhong-Huai Xiang, et al. "Topological and Functional Characterization of an Insect Gustatory Receptor". In: *PLoS ONE* 6.8 (Aug. 2011).

[73] Jae Young Kwon, Anupama Dahanukar, Linnea A. Weiss, and John R. Carlson. "The molecular basis of CO2 reception in Drosophila". In: *Proc Natl Acad Sci U S A* 104.9 (Feb. 2007).

[74] Mattias C Larsson, Ana I Domingos, Walton D Jones, M Eugenia Chiappe, Hubert Amrein, et al. "Or83b encodes a broadly expressed odorant receptor essential for Drosophila olfaction". In: *Neuron* 43.5 (Sept. 2004).

[75] Vincent Croset, Raphael Rytz, Scott F. Cummins, Aidan Budd, David Brawand, et al. "Ancient Protostome Origin of Chemosensory Ionotropic Glutamate Receptors and the Evolution of Insect Taste and Olfaction". In: *PLoS Genet* 6.8 (Aug. 2010).

[76] Philippe P Laissue and Leslie B Vosshall. "The olfactory sensory map in Drosophila". In: *Adv. Exp. Med. Biol.* 628 (2008).

[77] Merid N. Getahun, Dieter Wicher, Bill S. Hansson, and Shannon B. Olsson. "Temporal response dynamics of Drosophila olfactory sensory neurons depends on receptor type and response polarity". In: *Front Cell Neurosci* 6 (Nov. 2012).

[78] Merid N. Getahun, Shannon B. Olsson, Sofia Lavista-Llanos, Bill S. Hansson, and Dieter Wicher. "Insect Odorant Response Sensitivity Is Tuned by Metabotropically Autoregulated Olfactory Receptors". In: *PLoS One* 8.3 (Mar. 2013).

[79] J S Edwards. "The evolution of insect flight: implications for the evolution of the nervous system". In: *Brain Behav. Evol.* 50.1 (July 1997).

[80] A. Sanchez-Gracia, F. G. Vieira, and J. Rozas. "Molecular evolution of the major chemosensory gene families in insects". In: *Heredity* 103.3 (2009).

[81] C. S. McBride. "Rapid evolution of smell and taste receptor genes during host specialization in Drosophila sechellia". In: *Proc Natl Acad Sci USA* 104 (2007).

[82] A. Gardiner, D. Barker, R. K. Butlin, W. C. Jordan, and M. G. Ritchie. "Drosophila chemoreceptor gene evolution: selection, specialization and genome size". In: *Mol Ecol* 17.7 (2008).

[83] Adriana D. Briscoe, Aide Macias-Muñoz, Krzysztof M. Kozak, James R. Walters, Furong Yuan, et al. "Female Behaviour Drives Expression and Evolution of Gustatory Receptors in Butterflies". In: *PLoS Genet* 9.7 (July 2013).

[84] K. Tamura, S. Subramanian, and S. Kumar. "Temporal patterns of fruit fly (Drosophila) evolution revealed by mutation clocks". In: *Mol Biol Evol* 21 (2004).

[85] S. Foret and R. Maleszka. "Function and evolution of a gene family encoding odorant binding-like proteins in a social insect, the honey bee (Apis mellifera)". In: *Genome Res* 16 (2006).

[86] Alejandro Sánchez-Gracia and Julio Rozas. "Divergent evolution and molecular adaptation in the Drosophila odorant-binding protein family: inferences from sequence variation at the OS-E and OS-F genes". In: *BMC Evol. Biol.* 8 (2008).

[87] A. Ray, W. van der Goes van Naters, and J. R. Carlson. "A regulatory code for neuron-specific odor receptor expression". In: *PLoS Biol* 6.5 (2008).

[88] Shadi Jafari, Liza Alkhori, Alexander Schleiffer, Anna Brochtrup, Thomas Hummel, et al. "Combinatorial Activation and Repression by Seven Transcription Factors Specify Drosophila Odorant Receptor Expression". In: *PLoS Biol* 10.3 (Mar. 2012).

[89] S. Zhou, E. A. Stone, T. F. Mackay, and R. R. Anholt. "Plasticity of the chemoreceptor repertoire in Drosophila melanogaster". In: *PLoS Genet* 5.10 (2009).

[90] P. Wang, R. F. Lyman, T. F. Mackay, and R. R. Anholt. "Natural variation in odorant recognition among odorant-binding proteins in Drosophila melanogaster". In: *Genetics* 184.3 (2010).

[91] Mira V. Han, Gregg W. C. Thomas, Jose Lugo-Martinez, and Matthew W. Hahn. "Estimating Gene Gain and Loss Rates in the Presence of Error in Genome Assembly and Annotation Using CAFE 3". In: *Mol Biol Evol* 30.8 (Aug. 2013).

[92] Noah K Whiteman and Naomi E Pierce. "Delicious poison: genetics of Drosophila host plant preference". In: *Trends Ecol. Evol. (Amst.)* 23.9 (Sept. 2008).

[93] Zev Wisotsky, Adriana Medina, Erica Freeman, and Anupama Dahanukar. "Evolutionary differences in food preference rely on Gr64e, a receptor for glycerol". In: *Nat. Neurosci.* 14.12 (Dec. 2011).

[94] J. M. Ranz, D. Maurin, Y. S. Chan, M. von Grotthuss, L. W. Hillier, et al. "Principles of genome evolution in the Drosophila melanogaster species group". In: *PLoS Biol* 5.6 (2007).

[95] A. Bhutkar, S. W. Schaeffer, S. M. Russo, M. Xu, T. F. Smith, et al. "Chromosomal rearrangement inferred from comparisons of 12 Drosophila genomes". In: *Genetics* 179.3 (2008).

[96] J. M. Ranz, F. Casals, and A. Ruiz. "How malleable is the eukaryotic genome? Extreme rate of chromosomal rearrangement in the genus Drosophila". In: *Genome Res* 11.2 (2001).

[97] M. von Grotthuss, M. Ashburner, and J. M. Ranz. "Fragile regions and not functional constraints predominate in shaping gene organization in the genus Drosophila". In: *Genome Res* 20.8 (2010).

[98] F. G. Vieira, A. Sanchez-Gracia, and J. Rozas. "Comparative genomic analysis of the odorant-binding protein family in 12 Drosophila genomes: purifying selection and birth-and-death evolution". In: *Genome Biol* 8.11 (2007).

[99] P. G. Engstrom, S. J. Ho Sui, O. Drivenes, T. S. Becker, and B. Lenhard. "Genomic regulatory blocks underlie extensive microsynteny conservation in insects". In: *Genome Res* 17.12 (2007).

[100] H. Caron, B. van Schaik, M. van der Mee, F. Baas, G. Riggins, et al. "The human transcriptome map: clustering of highly expressed genes in chromosomal domains". In: *Science* 291.5507 (2001).

[101] M. J. Lercher, A. O. Urrutia, and L. D. Hurst. "Clustering of housekeeping genes provides a unified model of gene order in the human genome". In: *Nat Genet* 31.2 (2002).

[102] N. N. Batada and L. D. Hurst. "Evolution of chromosome organization driven by selection for reduced gene expression noise". In: *Nat Genet* 39.8 (2007).

[103] Z. Wang and J. Zhang. "Impact of gene expression noise on organismal fitness and the efficacy of natural selection". In: *Proc Natl Acad Sci U S A* 108.16 (2010).

[104] H. A. Wallace, M. P. Plata, H. J. Kang, M. Ross, and M. Labrador. "Chromatin insulators specifically associate with different levels of higher-order chromatin organization in Drosophila". In: *Chromosoma* 119.2 (2009).

[105] E. A. Rach, D. R. Winter, A. M. Benjamin, D. L. Corcoran, T. Ni, et al. "Transcription initiation patterns indicate divergent strategies for gene regulation at the chromatin level". In: *PLoS Genet* 7.1 (2011).

[106] X. Bao, W. Zhang, R. Krencik, H. Deng, Y. Wang, et al. "The JIL-1 kinase interacts with lamin Dm0 and regulates nuclear lamina morphology of Drosophila nurse cells". In: *J Cell Sci* 118.Pt 21 (2005).

[107] Helen Pickersgill, Bernike Kalverda, Elzo de Wit, Wendy Talhout, Maarten Fornerod, et al. "Characterization of the Drosophila melanogaster genome at the nuclear lamina". In: *Nat Genet* 38.9 (Sept. 2006).

[108] U. Rath, Y. Ding, H. Deng, H. Qi, X. Bao, et al. "The chromodomain protein, Chromator, interacts with JIL-1 kinase and regulates the structure of Drosophila polytene chromosomes". In: *J Cell Sci* 119.Pt 11 (2006).

[109] Jose M. Ranz, Carlos Diaz-Castillo, and Rita Petersen. "Conserved Gene Order at the Nuclear Periphery in Drosophila". In: *Molecular Biology and Evolution* 29.1 (2011).

[110] M. D. Adams, S. E. Celniker, R. A. Holt, C. A. Evans, J. D. Gocayne, et al. "The genome sequence of Drosophila melanogaster". In: *Science* 287.5461 (2000).

[111] S. Richards, Y. Liu, B. R. Bettencourt, P. Hradecky, S. Letovsky, et al. "Comparative genome sequencing of Drosophila pseudoobscura: chromosomal, gene, and cis-element evolution". In: *Genome Res* 15 (2005).

[112] S. Roy, J. Ernst, P. V. Kharchenko, P. Kheradpour, N. Negre, et al. "Identification of functional elements and regulatory circuits by Drosophila modENCODE". In: *Science* 330.6012 (2010).

[113] J. M. Comeron, R. Ratnappan, and S. Bailin. "The many landscapes of recombination in Drosophila melanogaster". In: *PLoS Genet* 8.10 (2012).

[114] T. F. Mackay, S. Richards, E. A. Stone, A. Barbadilla, J. F. Ayroles, et al. "The Drosophila melanogaster Genetic Reference Panel". In: *Nature* 482.7384 (2012).

[115] Julius Adler and Wung-Wai Tso. ""Decision"-Making in Bacteria: Chemotactic Response of Escherichia coli to Conflicting Stimuli". In: *Science* 184.4143 (1974), pp. 1292–1294.

[116] P. Librado and J. Rozas. "DnaSP v5: a software for comprehensive analysis of DNA polymorphism data". In: *Bioinformatics* 25.11 (2009).

[117] W. James Kent, Charles W. Sugnet, Terrence S. Furey, Krishna M. Roskin, Tom H. Pringle, et al. "The Human Genome Browser at UCSC". In: *Genome Research* 12.6 (2002), pp. 996–1006.

[118] Lincoln D. Stein, Christopher Mungall, ShengQiang Shu, Michael Caudy, Marco Mangone, et al. "The Generic Genome Browser: A Building Block for a Model Organism System Database". In: *Genome Research* 12.10 (2002), pp. 1599–1610.

[119] Stephen C J Parker, Loren Hansen, Hatice Ozel Abaan, Thomas D Tullius, and Elliott H Margulies. "Local DNA topography correlates with functional noncoding regions of the human genome". In: *Science* 324.5925 (Apr. 2009).

[120] Remo Rohs, Xiangshu Jin, Sean M West, Rohit Joshi, Barry Honig, et al. "Origins of specificity in protein-DNA recognition". In: *Annu. Rev. Biochem.* 79 (2010).

[121] M C King and A C Wilson. "Evolution at two levels in humans and chimpanzees". In: *Science* 188.4184 (Apr. 1975), pp. 107–116.

[122] Julio Rozas and Ricardo Rozas. "DnaSP, DNA sequence polymorphism: an interactive program for estimating population genetics parameters from DNA sequence data". In: *Computer applications in the biosciences: CABIOS* 11.6 (1995), pp. 621–625.

[123] Julio Rozas and Ricardo Rozas. "DnaSP version 2.0: a novel software package for extensive molecular population genetics analysis." In: *Computer applications in the biosciences: CABIOS* 13.3 (1997), pp. 307–311.

[124] Julio Rozas and Ricardo Rozas. "DnaSP version 3: an integrated program for molecular population genetics and molecular evolution analysis." In: *Bioinformatics* 15.2 (1999), pp. 174–175.

[125] Julio Rozas, Juan C. Sánchez-DelBarrio, Xavier Messeguer, and Ricardo Rozas. "DnaSP, DNA polymorphism analyses by the coalescent and other methods". In: *Bioinformatics* 19.18 (2003), pp. 2496–2497.

[126] D. Boffelli, J. McAuliffe, D. Ovcharenko, K. D. Lewis, I. Ovcharenko, et al. "Phylogenetic shadowing of primate sequences to find functional regions of the human genome". In: *Science* 299.5611 (2003).

[127] Jaime Huerta-Cepas, Hernán Dopazo, Joaquín Dopazo, and Toni Gabaldón. "The human phylome". In: *Genome Biology* 8.6 (June 2007), R109.

[128] L. Arvestad, A. C. Berglund, J. Lagergren, and B. Sennblad. "Bayesian gene/species tree reconciliation and orthology analysis using MCMC". In: *Bioinformatics* 19 Suppl 1 (2003).

[129] Bastien Boussau, Gergely J. Szöllősi, Laurent Duret, Manolo Gouy, Eric Tannier, et al. "Genome-scale coestimation of species and gene trees". In: *Genome Res.* 23.2 (Feb. 2013).

[130] Charles-Elie Rabier, Tram Ta, and Cécile Ané. "Detecting and Locating Whole Genome Duplications on a Phylogeny: A Probabilistic Approach". In: *Mol Biol Evol* (Dec. 2013), mst263.

[131] M. W. Hahn, M. V. Han, and S. G. Han. "Gene family evolution across 12 Drosophila genomes". In: *PLoS Genet* 11 (2007).

[132] Yan-Yan Wang, Bin Liu, Xin-Yu Zhang, Qi-Ming Zhou, Tao Zhang, et al. "Genome characteristics reveal the impact of lichenization on lichen-forming fungus Endocarpon pusillum Hedwig (Verrucariales, Ascomycota)". In: *BMC Genomics* 15.1 (Jan. 2014), p. 34.

[133] Jeffery P. Demuth, Tijl De Bie, Jason E. Stajich, Nello Cristianini, and Matthew W. Hahn. "The Evolution of Mammalian Gene Families". In: *PLoS ONE* 1.1 (Dec. 2006), e85.

[134] P. Librado, F. G. Vieira, and J. Rozas. "BadiRate: Estimating Family Turnover Rates by Likelihood-Based Methods". In: *Bioinformatics* 28.2 (2012).

[135] Jay Shendure and Hanlee Ji. "Next-generation DNA sequencing". In: *Nat Biotech* 26.10 (Oct. 2008), pp. 1135–1145.

[136] J. E. Pool, R. B. Corbett-Detig, R. P. Sugino, K. A. Stevens, C. M. Cardeno, et al. "Population Genomics of sub-saharan Drosophila melanogaster: African diversity and non-African admixture". In: *PLoS Genet* 8.12 (2013).

[137] Casey Bergman. *Release of 20 European Drosophila melanogaster genomes*. July 2012.

[138] M. Ramia, P. Librado, S. Casillas, J. Rozas, and A. Barbadilla. "PopDrowser: the Population Drosophila Browser". In: *Bioinformatics* 28.4 (2012).

[139] A. M. Boutanaev, A. I. Kalmykova, Y. Y. Shevelyov, and D. I. Nurminsky. "Large clusters of co-expressed genes in the Drosophila genome". In: *Nature* 420.6916 (2002).

[140] T. Itoh, K. Takemoto, H. Mori, and T. Gojobori. "Evolutionary instability of operon structures disclosed by sequence comparisons of complete microbial genomes". In: *Mol Biol Evol* 16.3 (1999).

[141] Hidemi Watanabe, Hirotada Mori, Takeshi Itoh, and Takashi Gojobori. "Genome plasticity as a paradigm of eubacteria evolution". In: *J Mol Evol* 44.1 (Jan. 1997), S57–S64.

[142] N. Luc, J. L. Risler, A. Bergeron, and M. Raffinot. "Gene teams: a new formalization of gene clusters for comparative genomics". In: *Comput Biol Chem* 27.1 (2003).

[143] P. Pevzner and G. Tesler. "Human and mouse genomic sequences reveal extensive breakpoint reuse in mammalian evolution". In: *Proc Natl Acad Sci U S A* 100.13 (2003).

[144] A. Ruiz-Herrera, J. Castresana, and T. J. Robinson. "Is mammalian chromosomal evolution driven by regions of genome fragility?" In: *Genome Biol* 7.12 (2006).

[145] X. Ling, X. He, and D. Xin. "Detecting gene clusters under evolutionary constraint in a large number of genomes". In: *Bioinformatics* 25.5 (2009).

[146] C. C. Weber and L. D. Hurst. "Support for multiple classes of local expression clusters in Drosophila melanogaster, but no evidence for gene order conservation". In: *Genome Biol* 12.3 (2011).

[147] F JACOB, D PERRIN, C SANCHEZ, and J MONOD. "[Operon: a group of genes with the expression coordinated by an operator]". In: *C. R. Hebd. Seances Acad. Sci.* 250 (Feb. 1960), pp. 1727–1729.

[148] Shigehiro Kuraku and Axel Meyer. "The evolution and maintenance of Hox gene clusters in vertebrates and the teleost-specific genome duplication". In: *Int. J. Dev. Biol.* 53.5-6 (2009), pp. 765–773.

[149] Simona Santini, Jeffrey L Boore, and Axel Meyer. "Evolutionary conservation of regulatory elements in vertebrate Hox gene clusters". In: *Genome Res.* 13.6A (June 2003), pp. 1111–1122.

[150] D. Noordermeer, M. Leleu, E. Splinter, J. Rougemont, W. De Laat, et al. "The dynamic architecture of Hox gene clusters". In: *Science* 334.6053 (2011).

[151] D. Dong, X. Shao, N. Deng, and Z. Zhang. "Gene expression variations are predictive for stochastic noise". In: *Nucleic Acids Res* 39.2 (2011).

[152] Z. Zhang, W. Qian, and J. Zhang. "Positive selection for elevated gene expression noise in yeast". In: *Mol Syst Biol* 5 (2009).

[153] P. Wang, R. F. Lyman, S. A. Shabalina, T. F. Mackay, and R. R. Anholt. "Association of polymorphisms in odorant-binding protein genes with variation in olfactory response to benzaldehyde in Drosophila". In: *Genetics* 177.3 (2007).

[154] S. Swarup, T. I. Williams, and R. R. Anholt. "Functional dissection of Odorant binding protein genes in Drosophila melanogaster". In: *Genes Brain Behav* 10.6 (2011).

[155] H. J. Muller. "Types of visible variations induced by X-rays inDrosophila". In: *Journ. of Gen.* 22.3 (July 1930), pp. 299–334.

[156] M. Gan, S. Moebus, H. Eggert, and H. Saumweber. "The Chriz-Z4 complex recruits JIL-1 to polytene chromosomes, a requirement for interband-specific phosphorylation of H3S10". In: *J Biosci* 36.3 (2011).

[157] W. A. Kellner, E. Ramos, K. Van Bortle, N. Takenaka, and V. G. Corces. "Genome-wide phosphoacetylation of histone H3 at Drosophila enhancers and promoters". In: *Genome Res* (2012).

[158] T. Rajala, A. Hakkinen, S. Healy, O. Yli-Harja, and A. S. Ribeiro. "Effects of transcriptional pausing on gene expression dynamics". In: *PLoS Comput Biol* 6.3 (2010).

[159] R. A. Hoskins, J. M. Landolin, J. B. Brown, J. E. Sandler, H. Takahashi, et al. "Genome-wide analysis of promoter architecture in Drosophila melanogaster". In: *Genome Res* 21.2 (2011).

[160] P. Duchen, D. Zivkovic, S. Hutter, W. Stephan, and S. Laurent. "Demographic inference reveals African and European admixture in the North American Drosophila melanogaster population". In: *Genetics* 193.1 (2012).

[161] Andreas Keller. "Drosophila melanogaster's history as a human commensal". In: *Curr. Biol.* 17.3 (Feb. 2007).

[162] F. Tajima. "The effect of change in population size on DNA polymorphism". In: *Genetics* 123.3 (1989).

[163] R. Egea, S. Casillas, and A. Barbadilla. "Standard and generalized McDonald-Kreitman test: a website to detect selection by comparing different classes of DNA sites". In: *Nucleic Acids Res* 36.Web Server issue (2008).

[164] J. Charlesworth and A. Eyre-Walker. "The McDonald-Kreitman test and slightly deleterious mutations". In: *Mol Biol Evol* 25.6 (2008).

[165] J. Parsch, Z. Zhang, and J. F. Baines. "The influence of demography and weak selection on the McDonald-Kreitman test: an empirical study in Drosophila". In: *Mol Biol Evol* 26.3 (2009).

[166] P. D. Keightley and A. Eyre-Walker. "Joint inference of the distribution of fitness effects of deleterious mutations and population demography based on nucleotide polymorphism frequencies". In: *Genetics* 177.4 (2007).

[167] J. C. Fay, G. J. Wyckoff, and C. I. Wu. "Positive and negative selection on the human genome". In: *Genetics* 158.3 (2001).

[168] T. I. Gossmann, P. D. Keightley, and A. Eyre-Walker. "The effect of variation in the effective population size on the rate of adaptive molecular evolution in eukaryotes". In: *Genome Biol Evol* 4.5 (2012).

[169] P. R. Haddrill, D. Bachtrog, and P. Andolfatto. "Positive and negative selection on noncoding DNA in Drosophila simulans". In: *Mol Biol Evol* 25.9 (2008).

[170] Mikhail Spivakov, Junaid Akhtar, Pouya Kheradpour, Kathryn Beal, Charles Girardot, et al. "Analysis of variation at transcription factor binding sites in Drosophila and humans". In: *Genome Biology* 13.9 (Sept. 2012), R49.

[171] R. K. Bradley, X. Y. Li, C. Trapnell, S. Davidson, L. Pachter, et al. "Binding site turnover produces pervasive quantitative changes in transcription factor binding between closely related Drosophila species". In: *PLoS Biol* 8.3 (2010).

[172] R. D. Dowell. "Transcription factor binding variation in the evolution of gene regulation". In: *Trends Genet* 26.11 (2010).

[173] D. Schmidt, M. D. Wilson, B. Ballester, P. C. Schwalie, G. D. Brown, et al. "Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding". In: *Science* 328.5981 (2010).

[174] M. Z. Ludwig, C. Bergman, N. H. Patel, and M. Kreitman. "Evidence for stabilizing selection in a eukaryotic enhancer element". In: *Nature* 403.6769 (2000).

[175] R. Durrett and D. Schmidt. "Waiting for two mutations: with applications to regulatory sequence evolution and the limits of Darwinian evolution". In: *Genetics* 180.3 (2008).

[176] B. Z. He, A. K. Holloway, S. J. Maerkl, and M. Kreitman. "Does positive selection drive transcription factor binding site turnover? A test with Drosophila cis-regulatory modules". In: *PLoS Genet* 7.4 (2011).

Anexo

A
El proyecto DGRP

The Drosophila melanogaster Genetic Reference Panel
Nature. 2012 Feb 8; 482 (7384): 173–178

ARTICLE

doi:10.1038/nature10811

The *Drosophila melanogaster* Genetic Reference Panel

Trudy F. C. Mackay[1]*, Stephen Richards[2]*, Eric A. Stone[1]*, Antonio Barbadilla[3]*, Julien F. Ayroles[1]†, Dianhui Zhu[2], Sònia Casillas[3]†, Yi Han[2], Michael M. Magwire[1], Julie M. Cridland[4], Mark F. Richardson[5], Robert R. H. Anholt[6], Maite Barrón[3], Crystal Bess[2], Kerstin Petra Blankenburg[2], Mary Anna Carbone[1], David Castellano[3], Lesley Chaboub[2], Laura Duncan[1], Zeke Harris[1], Mehwish Javaid[2], Joy Christina Jayaseelan[2], Shalini N. Jhangiani[2], Katherine W. Jordan[1], Fremiet Lara[2], Faye Lawrence[1], Sandra L. Lee[2], Pablo Librado[7], Raquel S. Linheiro[5], Richard F. Lyman[1], Aaron J. Mackey[8], Mala Munidasa[2], Donna Marie Muzny[2], Lynne Nazareth[2], Irene Newsham[2], Lora Perales[2], Ling-Ling Pu[2], Carson Qu[2], Miquel Ràmia[3], Jeffrey G. Reid[2], Stephanie M. Rollmann[1]†, Julio Rozas[7], Nehad Saada[2], Lavanya Turlapati[1], Kim C. Worley[2], Yuan-Qing Wu[2], Akihiko Yamamoto[1], Yiming Zhu[2], Casey M. Bergman[5], Kevin R. Thornton[4], David Mittelman[9] & Richard A. Gibbs[2]

A major challenge of biology is understanding the relationship between molecular genetic variation and variation in quantitative traits, including fitness. This relationship determines our ability to predict phenotypes from genotypes and to understand how evolutionary forces shape variation within and between species. Previous efforts to dissect the genotype–phenotype map were based on incomplete genotypic information. Here, we describe the *Drosophila melanogaster* Genetic Reference Panel (DGRP), a community resource for analysis of population genomics and quantitative traits. The DGRP consists of fully sequenced inbred lines derived from a natural population. Population genomic analyses reveal reduced polymorphism in centromeric autosomal regions and the *X* chromosome, evidence for positive and negative selection, and rapid evolution of the *X* chromosome. Many variants in novel genes, most at low frequency, are associated with quantitative traits and explain a large fraction of the phenotypic variance. The DGRP facilitates genotype–phenotype mapping using the power of *Drosophila* genetics.

Understanding how molecular variation maps to phenotypic variation for quantitative traits is central for understanding evolution, animal and plant breeding, and personalized medicine[1,2]. The principles of mapping quantitative trait loci (QTLs) by linkage to, or association with, marker loci are conceptually simple[1,2]. However, we have not yet achieved our goal of explaining genetic variation for quantitative traits in terms of the underlying genes; additive, epistatic and pleiotropic effects as well as phenotypic plasticity of segregating alleles; and the molecular nature, population frequency and evolutionary dynamics of causal variants. Efforts to dissect the genotype–phenotype map in model organisms[3,4] and humans[5–7] have revealed unexpected complexities, implicating many, novel loci, pervasive pleiotropy, and context-dependent effects.

Model organism reference populations of inbred strains that can be shared among laboratories studying diverse phenotypes, and for which environmental conditions can be controlled and manipulated, greatly facilitate efforts to dissect the genetic architecture of quantitative traits[3,4]. Measuring many individuals of the same homozygous genotype increases the accuracy of the estimates of genotypic value[1] and the power to detect variants, and genotypes of molecular markers need only be obtained once. We constructed the *Drosophila melanogaster* Genetic Reference Panel (DGRP) as such a community resource. Unlike previous populations of recombinant inbred lines derived from limited samples of genetic variation, the DGRP consists of 192 inbred strains derived from a single outbred population. The DGRP contains a representative sample of naturally segregating genetic variation, has an ultra-fine-grained recombination map suitable for precise localization of causal variants, and has almost complete euchromatic sequence information.

Here, we describe molecular and phenotypic variation in 168 re-sequenced lines comprising Freeze 1.0 of the DGRP, population genomic inferences of patterns of polymorphism and divergence and their correlation with genomic features, local recombination rate and selection acting on this population, genome-wide association mapping analyses for three quantitative traits, and tools facilitating the use of this resource.

Molecular variation in the DGRP

We constructed the DGRP by collecting mated females from the Raleigh, North Carolina, USA, population, followed by 20 generations of full-sibling inbreeding of their progeny. We sequenced 168 DGRP lines using a combination of Illumina and 454 sequencing technology: 29 of the lines were sequenced using both platforms, 129 lines have only Illumina sequence, and 10 lines have only 454 sequence. We mapped sequence reads to the *D. melanogaster* reference genome, re-calibrated base quality scores, and locally re-aligned Illumina reads. Mean sequence coverage was 21.4× per line for Illumina sequences and 12.1× per line for 454 sequences (Supplementary

[1]Department of Genetics, North Carolina State University, Raleigh, North Carolina 27695, USA. [2]Human Genome Sequencing Center, Baylor College of Medicine, One Baylor Plaza, Houston, Texas 77030 USA. [3]Genomics, Bioinformatics and Evolution Group, Institut de Biotecnologia i de Biomedicina - IBB/Department of Genetics and Microbiology, Campus Universitat Autònoma de Barcelona, 08193 Bellaterra, Spain. [4]Department of Ecology and Evolutionary Biology, University of California - Irvine, Irvine, California 92697, USA. [5]Faculty of Life Sciences, University of Manchester, Manchester M13 9PT, UK. [6]Department of Biology, North Carolina State University, Raleigh, North Carolina 27695, USA. [7]Molecular Evolutionary Genetics Group, Department of Genetics, Faculty of Biology, Universitat de Barcelona, Diagonal 645, 08028 Barcelona, Spain. [8]Center for Public Health Genomics, University of Virginia, PO Box 800717, Charlottesville, Virginia 22908, USA. [9]Virginia Bioinformatics Institute and Department of Biological Sciences, Virginia Tech, Blacksburg, Virginia 24061, USA. †Present addresses: FAS Society of Fellows, Harvard University, 78 Mt Auburn Street, Cambridge, Massachusetts 02138, USA (J.F.A.); Functional Comparative Genomics Group, Institut de Biotecnologia i de Biomedicina - IBB, Campus Universitat Autònoma de Barcelona, 08193 Bellaterra, Spain (S.C.); Department of Biological Sciences, University of Cincinnati, Cincinnati, Ohio 45221, USA (S.M.R.).
*These authors contributed equally to this work.

Table 1). On average, we assayed 113.5 megabases (94.25%) of the euchromatic reference sequence with ~22,000 read mapping gaps per line (Supplementary Table 2). We called 4,672,297 single nucleotide polymorphisms (SNPs) using the Joint Genotyper for Inbred Lines (JGIL; E.A.S., personal communication), which takes into account coverage and quality sequencing statistics, and expected allele frequencies after 20 generations of inbreeding from an outbred population initially in Hardy–Weinberg equilibrium. In cases where base calls were made by both technologies, concordance was 99.36% (Supplementary Table 3).

The SNP site frequency distribution (Fig. 1a) is characterized by a majority of low frequency variants. The numbers of SNPs vary by chromosome and site class (Fig. 1b). Linkage disequilibrium[8] decays to $r^2 = 0.2$ on average within 10 base pairs on autosomes and 30 base pairs on the *X* chromosome (Fig. 1c and Supplementary Fig. 1). This difference is expected because the population size of the *X* chromosome is three quarters that of autosomes, and the *X* chromosome can experience greater purifying selection because of exposure of deleterious recessive alleles in hemizygous males. There is little evidence of global population structure in the DGRP (Fig. 1d and Supplementary Fig. 2). The rapid decline in linkage disequilibrium locally and lack of global population structure are favourable for genome-wide association mapping.

Not all SNPs are fixed within individual DGRP lines (Supplementary Table 4). The expected inbreeding coefficient (F) after 20 generations of full-sibling inbreeding[1] is $F = 0.986$; therefore, we expect some SNPs to remain segregating by chance. Segregating SNPs can also arise from new mutations, or if natural selection opposes inbreeding, due to true overdominance for fitness at individual loci or associative overdominance due to complementary deleterious alleles that are closely linked or in segregating inversions.

We identified 390,873 microsatellite loci, 105,799 of which were polymorphic (Supplementary Table 5); 36,810 transposable element insertion sites and 197,402 total insertions (Supplementary Table 6). On average, each line contained 1,175 transposable element insertions (Supplementary Table 6), although most transposable element insertion sites (25,562) were present in only one line (Supplementary Table 7). We identified 149 transposable element families. The number of copies per family varied greatly from an average of 315.7 *INE-1* elements per line to an average of 0.003 *Gandalf-Dkoe-like* elements per line (Supplementary Table 8).

Wolbachia pipientis is a maternally inherited bacterium found in insects, including *Drosophila*, and can affect reproduction[9]. We assessed *Wolbachia* infection status in the DGRP lines to account for it in analyses of genotype–phenotype associations, and found 51.2% of lines harbouring sufficient *Wolbachia* DNA to imply infection (Supplementary Table 9).

Polymorphism and divergence

We used the DGRP Illumina sequence data and genome sequences from *Drosophila simulans* and *Drosophila yakuba*[10] to perform genome-wide analyses of polymorphism and divergence, assess the association of these parameters with genomic features and the recombination landscape, and infer the historical action of selection on a much larger scale than had been possible previously[11–16]. We computed polymorphism (π and θ, refs 17 and 18) and divergence (k, ref. 19) for the whole genome, by chromosome arm (*X*, *2L*, *2R*, *3L*, *3R*), by chromosome region (three regions of equal size in Mb — telomeric, middle and centromeric), in 50-kbp non-overlapping windows, and by site class (synonymous and non-synonymous sites within coding sequences, and intronic, untranslated region (UTR) and intergenic sites) (Supplementary Tables 10 and 11).

Averaged over the entire genome, $\pi = 0.0056$ and $\theta = 0.0067$, similar to previous estimates from North American populations[16,20]. Average polymorphism on the *X* chromosome ($\pi_X = 0.0040$) is reduced relative to the autosomes ($\pi_A = 0.0060$) (X/A ratio = 0.67, Wilcoxon test $P = 0$), even after correcting for the X/A effective population size ($X_{4/3} = 0.0054$, Wilcoxon test $P < 0.00002$; Supplementary Table 10). Autosomal nucleotide diversity is reduced on average 2.4-fold in centromeric regions relative to non-centromeric regions, and at the telomeres (Fig. 2a and Supplementary Table 10), whereas diversity is relatively constant along the *X* chromosome. Thus, $\pi_X > \pi_A$ in centromeric regions, but $\pi_A > \pi_X$ in other chromosomal regions (Fig. 2a and Supplementary Table 10).

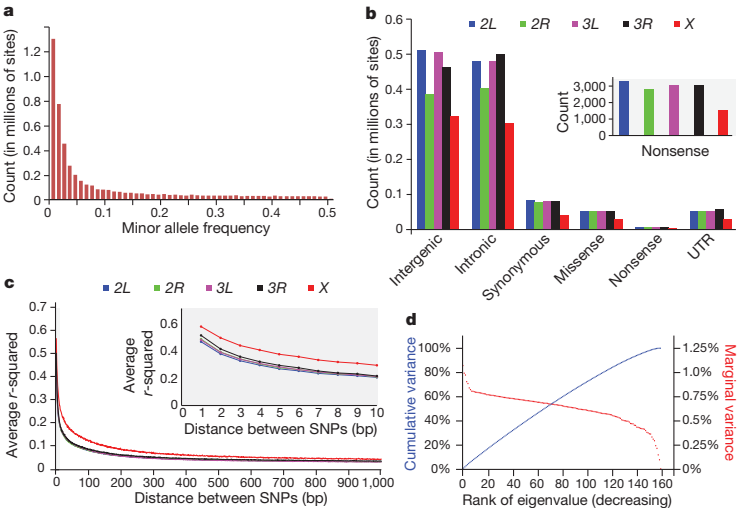


Figure 1 | SNP variation in the DGRP lines. **a**, Site frequency spectrum. **b**, Numbers of SNPs per site class. **c**, Decay of linkage disequilibrium (r^2) with physical distance for the five major chromosome arms. **d**, Lack of population structure. The red curve depicts the ranked eigenvalues of the genetic covariance matrix in decreasing order with respect to the marginal variance explained; the blue curve shows their cumulative sum as a fraction of the total with respect to cumulative variance explained. The partitioning of total genetic variance is balanced among the eigenvectors. The principal eigenvector explains < 1.1% of the total genetic variance.

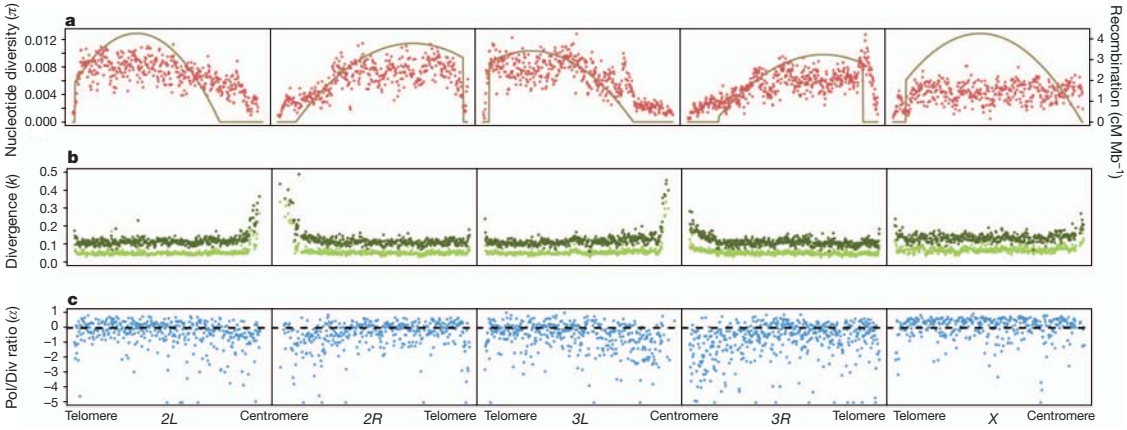


Figure 2 | Pattern of polymorphism, divergence, α and recombination rate along chromosome arms in non-overlapping 50-kbp windows. **a**, Nucleotide polymorphism (π). The solid curves give the recombination rate (cM Mb^{-1}). **b**, Divergence (k) for *D. simulans* (light green) and *D. yakuba* (dark green). **c**, Polymorphism to divergence ratio (Pol/Div), estimated as $1 - [(\pi_{0\text{-fold}}/\pi_{4\text{-fold}})/(k_{0\text{-fold}}/k_{4\text{-fold}})]$. An excess of 0-fold divergence relative to polymorphism $(k_{0\text{-fold}}/k_{4\text{-fold}}) > (\pi_{0\text{-fold}}/\pi_{4\text{-fold}})$ is interpreted as adaptive fixation whereas an excess of 0-fold polymorphism relative to divergence $(\pi_{0\text{-fold}}/\pi_{4\text{-fold}}) > (k_{0\text{-fold}}/k_{4\text{-fold}})$ indicates that weakly deleterious or nearly neutral mutations are segregating in the population.

Genes on the *X* chromosome evolve faster ($k_X = 0.140$) than autosomal genes ($k_A = 0.126$) (*X/A* ratio = 1.131, Wilcoxon test $P = 0$) (Fig. 2b and Supplementary Table 10). Divergence is more uniform (coefficient of variation $(CV)_k = 0.2841$) across chromosome arms than is polymorphism ($CV_\pi = 0.4265$). The peaks of divergence near the centromeres could be attributable to the reduced quality of alignments in these regions. Patterns of divergence are similar regardless of the outgroup species used (Fig. 2b and Supplementary Table 11).

The pattern of polymorphism and divergence by site class is consistent within and among chromosomes ($\pi_{k_{\text{Synonymous}}} > \pi_{k_{\text{Intron}}} > \pi_{k_{\text{Intergenic}}} > \pi_{k_{\text{UTR}}} > \pi_{k_{\text{Non-synonymous}}}$), in agreement with previous studies on smaller data sets[12,15] (Supplementary Figs 3 and 4 and Supplementary Table 11). Polymorphism levels between synonymous and non-synonymous sites differ by an order of magnitude. Variation and divergence patterns within the site classes generally follow the same patterns observed overall, with reduced polymorphism for all site classes on the *X* chromosome relative to autosomes, increased *X* chromosome divergence relative to autosomes for all but synonymous sites, decreased polymorphism in centromeric regions, and greater variation among regions and arms for polymorphism than for divergence. Other diversity measures and more detailed patterns at different window-sizes for each chromosome arm can be accessed from the Population *Drosophila* Browser (popDrowser) (Table 1 and Methods).

Recombination landscape

Evolutionary models of hitchhiking and background selection[21,22] predict a positive correlation between polymorphism and recombination rate. This expectation is realized in regions where recombination is less than 2 cM Mb^{-1} (Spearman's $\rho = 0.471$, $P = 0$), but recombination and polymorphism are independent in regions where recombination exceeds 2 cM Mb^{-1} (Spearman's $\rho = -0.0044$, $P = 0.987$) (Fig. 2a and Supplementary Table 12). The average rate of recombination of the *X* chromosome (2.9 cM Mb^{-1}) is greater than that of autosomes (2.1 cM Mb^{-1}), which may account for the low overall *X*-linked correlation between recombination rate and π. The lack of correlation between recombination and divergence (Supplementary Table 12) excludes mutation associated with recombination as the cause of the correlation. We assessed the independent effects of recombination rate, divergence, chromosome region and gene density on nucleotide variation of autosomes and the *X* chromosome (Supplementary Table 13). Recombination is the major predictor of polymorphism on the *X* chromosome and autosomes; however, the significant effect of autosomal chromosome region remains after accounting for variation in recombination rates between centromeric and non-centromeric regions.

Selection regimes

We used the standard[23] and generalized[12,24,25] McDonald Kreitman tests (MKT) to scan the genome for evidence of selection. These tests

Table 1 | Community resources

Resource	Location
DGRP lines	Bloomington *Drosophila* Stock Center http://flystocks.bio.indiana.edu/Browse/RAL.php
Sequences	Baylor College of Medicine Human Genome Sequencing Center http://www.hgsc.bcm.tmc.edu/project-species-i-DGRP_lines.hgsc National Center for Biotechnology Information Short Read Archive http://www.ncbi.nlm.nih.gov/sra?term=DGRP Mackay Laboratory http://dgrp.gnets.ncsu.edu/
Read alignments	Baylor College of Medicine Human Genome Sequencing Center http://www.hgsc.bcm.tmc.edu/projects/dgrp/
SNPs	Baylor College of Medicine Human Genome Sequencing Center http://www.hgsc.bcm.tmc.edu/projects/dgrp/freeze1_July_2010/snp_calls/ National Center for Biotechnology Information dbSNP http://www.ncbi.nlm.nih.gov/SNP/snp_viewBatch.cgi?sbid=1052186 Mackay Laboratory http://dgrp.gnets.ncsu.edu/
Microsatellites	Baylor College of Medicine Human Genome Sequencing Center http://www.hgsc.bcm.tmc.edu/projects/dgrp/freeze1_July_2010/microsat_calls/ Mittelman Laboratory http://genome.vbi.vt.edu/public/DGRP/
Transposable elements	Mackay Laboratory http://dgrp.gnets.ncsu.edu/
Molecular population genomics	PopDrowser http://popdrowser.uab.cat
Phenotypes	Mackay Laboratory http://dgrp.gnets.ncsu.edu/
Genome-wide association analysis	Mackay Laboratory http://dgrp.gnets.ncsu.edu/

compare the ratio of polymorphism at a selected site with that of a neutral site to the ratio of divergence at a selected site to divergence at a neutral site. The standard MKT is applied to coding sequences, and synonymous and non-synonymous sites are used as putative neutral and selected sites, respectively. The generalized MKT is applied to non-coding sequences and uses fourfold degenerate sites as neutral sites. Using polymorphism and divergence data avoids confounding inference of selection with mutation rate differences, and restricting the tests to closely linked sites controls for shared evolutionary history[26–28]. We infer adaptive divergence when there is an excess of divergence relative to polymorphism, and segregation of slightly deleterious mutations when there is an excess of polymorphism over divergence. Estimates of α, the proportion of adaptive divergence, are biased downwards by low frequency, slightly deleterious mutations[29,30]. Rather than eliminate low frequency variants[31], we incorporated information on the site frequency distribution to the MKT test framework to obtain estimates of the proportion of sites that are strongly deleterious (d), weakly deleterious (b), neutral (f) and recently neutral (γ) at segregating sites, as well as unbiased estimates of α (Supplementary Methods).

Deleterious and neutral sites

Averaged over the entire genome, we infer that 58.5% of the segregating sites are neutral or nearly neutral, 1.9% are weakly deleterious and 39.6% are strongly deleterious. However, these proportions vary between the *X* chromosome and autosomes, site classes and chromosome regions (Supplementary Tables 14–16 and Fig. 3). Non-synonymous sites are the most constrained ($d = 77.6\%$), whereas in non-coding sites d ranges from 29.1% in 5′ UTRs to 41.3% in 3′ intergenic regions. The inferred pattern of selection differs between autosomal centromeric and non-centromeric regions: d is reduced and f is increased in centromeric regions for all site categories (Fig. 3). We observe an excess of polymorphism relative to divergence in autosomal centromeric regions, even after correcting for weakly deleterious mutations, implying a relaxation of selection from the time of separation of *D. melanogaster* and *D. yakuba*. Because selection coefficients depend on the effective population size[32] (N_e), this could occur if the recombination rate has specifically diminished in centromeric regions during the divergence between *D. melanogaster* and *D. yakuba*; or with an overall reduction of N_e associated with the colonization of North American habitats[33,34]. In the latter case, we expect a genome-wide signature of an excess of low-frequency polymorphisms and of polymorphism relative to divergence, exacerbated in regions of low recombination. We indeed find an excess of low-frequency polymorphism relative to neutral expectation as indicated by the negative estimates of Tajima's D statistic[35] ($D = -0.686$ averaged over the whole genome and $D = -0.997$ in autosomal centromeric regions). In contrast, the *X* chromosome does not show a differential pattern of selection in the centromeric region, has a lower fraction of relaxation of selection, fewer neutral alleles, and a higher percentage of strongly deleterious alleles for all site classes and regions (Fig. 3 and Supplementary Tables 14–16).

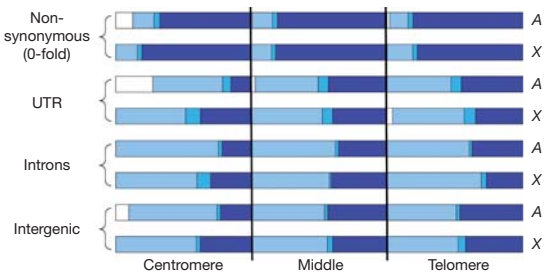


Figure 3 | The fraction of alleles segregating under different selection regimes by site class and chromosome region, for the autosomes (*A*) and the X chromosome (*X*). The selection regimes are strongly deleterious (d, dark blue), weakly deleterious (b, blue), recently neutral (γ, white) and old neutral ($f - \gamma$, light blue). Each chromosome arm has been divided in three regions of equal size (in Mb): centromere, middle and telomere.

Transposable element insertions are thought to be largely deleterious. There are more singleton insertions in regions of high recombination (≥ 2 cM Mb^{-1}) and more insertions shared in multiple lines in regions of low recombination (< 2 cM Mb^{-1}) (Fisher's exact test $P = 0$), and comparison of observed and expected site occupancy spectra reveals an excess of singleton insertions ($P = 0$, Supplementary Fig. 5).

Adaptive fixation

We find substantial evidence for positive selection in autosomal non-centromeric regions and the *X* chromosome (Fig. 2c and Supplementary Tables 15 and 17). We estimated α by aggregating all sites in each region analysed to avoid underestimation by averaging across genes[36] in comparisons of chromosomes, regions and site classes. We also computed the direction of selection, DoS[37], which is positive with adaptive selection, zero under neutrality and negative when weakly deleterious or new nearly neutral mutations are segregating. Estimates of α from the standard and generalized MKT indicate that on average 25.2% of the fixed sites between *D. melanogaster* and *D. yakuba* are adaptive, ranging from 30% in introns to 7% in UTR sites (Supplementary Fig. 6). Estimates of DoS and α are negative for non-synonymous and UTR sites in the autosomal centromeres, consistent with underestimating the fraction of adaptive substitutions in regions of low recombination because weakly deleterious or nearly neutral mutations are more common than adaptive fixations. The majority of adaptive fixation on autosomes occurs in non-centromeric regions (Fig. 2c). We find over four times as many adaptive fixations on the *X* chromosome relative to autosomes. The pattern holds for all site classes, in particular non-synonymous sites and UTRs, as well as individual genes, and is not solely due to the autosomal centromeric effect (Supplementary Table 15 and Supplementary Figs 6 and 7). Finally, when we consider DoS in recombination environments above and below 2 cM Mb^{-1}, we find greater adaptive propensity in genes whose recombination context is ≥ 2 cM Mb^{-1} (Wilcoxon test, $P = 0$; Supplementary Fig. 8).

To understand the global patterns of divergence and constraint across functional classes of genes, we examined the distributions of ω (d_N/d_S, the ratio of non-synonymous to synonymous divergence) and DoS across gene ontology (GO) categories. The 4.9% GO categories with significantly elevated DoS include the biological process categories of behaviour, developmental process involved in reproduction, reproduction and ion transport (Supplementary Table 18). Recombination context is the major determinant of variation in DoS (Supplementary Table 19) whereas GO category is as important as recombinational context for predicting variation in ω (Supplementary Table 19).

GO categories enriched for positive DoS values differ from those associated with high values of ω (Supplementary Table 18), indicating that positive selection does not occur necessarily on genes with high ω values. If adaptive substitutions are common, high values of ω reflect the joint contributions of neutral and adaptive substitutions. Further, equating high constraint (low ω) with functional importance overlooks the functional role of adaptive changes[15]. Unlike ω, DoS takes into account the constraints inferred from the current polymorphism, distinguishing negative, neutral and adaptive selection.

Genome-wide genotype–phenotype associations

We measured resistance to starvation stress, chill coma recovery time and startle response[38] in the DGRP. We found considerable genetic variation for all traits, with high broad sense heritabilities. We also found variation in sex dimorphism for starvation resistance and chill

coma recovery with cross-sex genetic correlations significantly different from unity (Supplementary Tables 20–22).

We performed genome-wide association analyses for these traits, using the 2,490,165 SNPs and 77,756 microsatellites for which the minor allele was represented in four or more lines, using single-locus analyses pooled across sexes and separately for males and females. At $P < 10^{-5}$ ($P < 10^{-6}$), we find 203 (32) SNPs and 2 (0) microsatellites associated with starvation resistance; 90 (7) SNPs and 4 (2) microsatellites associated with startle response; and 235 (45) SNPs and 5 (3) microsatellites associated with chill coma recovery time (Fig. 4a, Supplementary Fig. 9 and Supplementary Tables 23 and 24). The minor allele frequencies for most of the associated SNPs are low, and there is an inverse relationship between effect sizes and minor allele frequency (Supplementary Fig. 10).

The DGRP is a powerful tool for rapidly reducing the search space for molecular variants affecting quantitative traits from the entire genome to candidate polymorphisms and genes. Although we cannot infer which of these polymorphisms are causal due to linkage disequilibrium between SNPs in close physical proximity as well as occasional spurious long range linkage disequilibrium (Fig. 4a and Supplementary Fig. 9), the candidate gene lists are likely to be enriched for causal variants. The majority of associations are in computationally predicted genes or genes with annotated functions not obviously associated with the three traits. However, genes previously associated with startle response[39] (*Sema-1a* and *Eip75B*) and starvation resistance[40] (*pnt*) were identified in this study; and a SNP in *CG3213*, previously identified in a *Drosophila* obesity screen[41], is associated with variation in starvation resistance. Several genes associated with quantitative traits are rapidly evolving (*psq*, *Egfr*; Supplementary Tables 17 and 23) or are plausible candidates based on SNP or gene ontology annotations (Supplementary Table 23).

Predicting phenotypes from genotypes

We used regression models to predict trait phenotypes from SNP genotypes and estimate the total variance explained by SNPs. The latter cannot be done by summing the individual contributions of the single marker effects because markers are not completely independent, and estimates of effects of single markers are biased when more than one locus affecting the trait segregates in the population. We derived gene-centred multiple regression models to estimate the effects of multiple SNPs simultaneously. In all cases 6–10 SNPs explain from 51–72% of the phenotypic variance and 65–90% of the genetic variance (Supplementary Tables 25 and 26 and Supplementary Figs 11–13). We also derived partial least square regression models using all SNPs for which the single marker effect was significant at $P < 10^{-5}$. These models explain 72–85% of the phenotypic variance (Fig. 4b, c and Supplementary Fig. 14).

Discussion

The DGRP lines, sequences, variant calls, phenotypes and web tools for molecular population genomics and genome-wide association analysis are publicly available (Table 1). The DGRP lines contain at least 4,672,297 SNPs, 105,799 polymorphic microsatellites and 36,810 transposable elements, as well as insertion/deletion events and copy number variants and are a valuable resource for understanding the genetic architecture of quantitative traits of ecological and evolutionary relevance as well as *Drosophila* models of human quantitative traits. These novel mutations have survived the sieve of natural selection and will enhance the functional annotation of the *Drosophila* genome, complementing the *Drosophila* Gene Disruption Project[42] and the *Drosophila* modENCODE project[43].

Genome-wide molecular population genetic analyses show that patterns of polymorphism, but not divergence, differ by autosomal chromosome region, and between the *X* chromosome and autosomes. Polymorphism is lower in autosomal centromeric than non-centromeric regions, but not for the *X* chromosome. We propose that the correlation of polymorphism with recombination in regions where recombination is < 2 cM Mb^{-1} is due to the reduced effective population size in regions of low recombination[8]. Selection is less efficient in regions of low recombination[32], consistent with our observation that the fraction of strongly deleterious mutations and positively selected sites are reduced in these regions.

All molecular population genomic analyses support the 'faster *X*' hypothesis[44]. Relative to the autosomes, the *X* chromosome shows lower polymorphism, faster rates of molecular evolution, a higher percentage of gene regions undergoing adaptive evolution, a higher fraction of strongly deleterious sites, and a lower level of weak negative selection and relaxation of selection. New *X*-linked mutations are directly exposed to selection each generation in hemizygous males, and the *X* chromosome has greater recombination than autosomes[44]; both of these factors could contribute to this observation.

Genome-wide association analyses of three fitness-related quantitative traits reveal hundreds of novel candidate genes, highlighting our ignorance of the genetic basis of complex traits. Most variants associated with the traits are at low frequency, and there is an inverse relationship between frequency and effect. Given that low-frequency alleles are likely to be deleterious for traits under directional or stabilizing selection, these results are consistent with the mutation–selection balance hypothesis[1] for the maintenance of quantitative genetic variation. Regression models incorporating significant SNPs

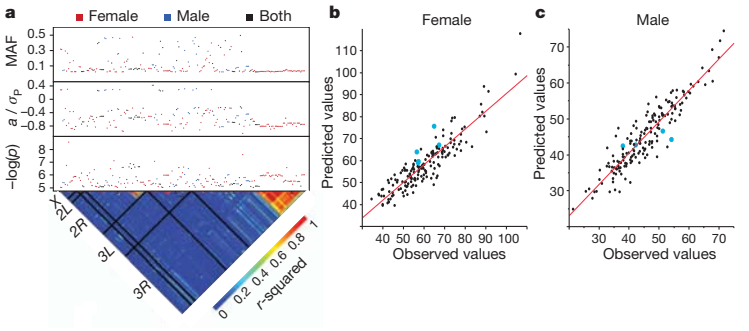


Figure 4 | Genotype–phenotype associations for starvation resistance. **a**, Genome-wide association results for significant SNPs. The lower triangle depicts linkage disequilibrium (r^2) among SNPs, with the five major chromosome arms demarcated by black lines. The upper panels give the significance threshold ($-\log(p)$, uncorrected for multiple tests), the effect in phenotypic standard deviation units, and the minor allele frequency (MAF). **b**, **c**, Partial least squares regressions of phenotypes predicted using SNP data on observed phenotypes. The blue dots represent the predicted and observed phenotypes of lines that were not included in the initial study. **b**, Females ($r^2 = 0.81$); **c**, males ($r^2 = 0.85$).

explain most of the phenotypic variance of the traits, in contrast with human association studies, where significant SNPs have tiny effects and together explain a small fraction of the total phenotypic variance[7]. If the genetic architecture of human complex traits is also dominated by low-frequency causal alleles, we expect estimates of effect size based on linkage disequilibrium with common variants to be strongly biased downwards.

In the future, the full power of *Drosophila* genetics can be applied to validating marker-trait associations: mutations, RNA interference constructs and quantitative trait loci mapping populations. The DGRP is an ideal resource for systems genetics analyses of the relationship between molecular variation, causal molecular networks and genetic variation for complex traits[4,38,45], and will anchor evolutionary studies in comparison with sequenced *Drosophila* species to assess to what extent variation within a species corresponds to variation among species.

METHODS SUMMARY

The full Methods are in the Supplementary Information. Information on sequencing and bioinformatics includes methods for DNA isolation; library construction and genomic sequencing; sequence read alignment; SNP, microsatellite and transposable element identification; genotypes for assurance of sample identity; and *Wolbachia* detection. Methods for molecular population genomics analysis include details of recombination estimates; diversity measures, linkage disequilibrium and neutrality tests; software used for population genomic analysis; data visualization (popDrowser); standard and generalized McDonald–Kreitman tests, statistical analysis methods; quality assessment and data filtering; and gene ontology analyses. Methods for quantitative genetic analyses include phenotype measures, quantitative genetic analyses of phenotypes, statistical analyses of genotype–phenotype associations and predictive models, and a web-based association analysis pipeline.

Received 13 July; accepted 21 December 2011.

1. Falconer, D. S. & Mackay, T. F. C. *Introduction to Quantitative Genetics* 4th edn (Longman, 1996).
2. Lynch, M. & Walsh, B. *Genetics and Analysis of Quantitative Traits* (Sinauer Associates, 1998).
3. Flint, J. & Mackay, T. F. C. Genetic architecture of quantitative traits in flies, mice and humans. *Genome Res.* **19,** 723–733 (2009).
4. Mackay, T. F. C., Stone, E. A. & Ayroles, J. F. The genetics of quantitative traits: challenges and prospects. *Nature Rev. Genet.* **10,** 565–577 (2009).
5. Altshuler, D., Daly, M. J. & Lander, E. S. Genetic mapping in human disease. *Science* **322,** 881–888 (2008).
6. Donnelly, P. Progress and challenges in genome-wide association studies in humans. *Nature* **456,** 728–731 (2008).
7. Manolio, T. A. *et al.* Finding the missing heritability of complex diseases. *Nature* **461,** 747–753 (2009).
8. Hill, W. G. & Robertson, A. The effect of linkage on limits to artificial selection. *Genet. Res.* **8,** 269–294 (1966).
9. Werren, J. H. Biology of *Wolbachia*. *Annu. Rev. Entomol.* **42,** 587–609 (1997).
10. Clark, A. G. *et al.* Evolution of genes and genomes on the *Drosophila* phylogeny. *Nature* **450,** 203–218 (2007).
11. Smith, N. G. & Eyre-Walker, A. Adaptive protein evolution in *Drosophila*. *Nature* **415,** 1022–1024 (2002).
12. Andolfatto, P. Adaptive evolution of non-coding DNA in *Drosophila*. *Nature* **437,** 1149–1152 (2005).
13. Presgraves, D. C. Recombination enhances protein adaptation in *Drosophila melanogaster*. *Curr. Biol.* **15,** 1651–1656 (2005).
14. Casillas, S., Barbadilla, A. & Bergman, C. Purifying selection maintains highly conserved noncoding sequences in *Drosophila*. *Mol. Biol. Evol.* **24,** 2222–2234 (2007).
15. Sella, G. *et al.* Pervasive natural selection in the *Drosophila* genome? *PLoS Genet.* **5,** e1000495 (2009).
16. Sackton, T. B. *et al.* Population genomic inferences from sparse high-throughput sequencing of two populations of *Drosophila melanogaster*. *Genome Biol. Evol.* **1,** 449–465 (2009).
17. Nei, M. *Molecular Evolutionary Genetics* (Columbia Univ. Press, 1987).
18. Watterson, G. A. On the number of segregating sites in genetical models without recombination. *Theor. Popul. Biol.* **7,** 256–276 (1975).
19. Jukes, T. H. & Cantor, C. R. in *Mammalian Protein Metabolism* vol. 3 (eds Munro, H. N. & Allison, J. B.) 21–132 (Academic Press, 1969).
20. Andolfatto, P. & Przeworski, M. Regions of lower crossing over harbor more rare variants in African *Drosophila melanogaster*. *Genetics* **158,** 657–665 (2001).
21. Begun, D. J. & Aquadro, C. F. Levels of naturally occurring DNA polymorphism correlate with recombination rates in *D. melanogaster*. *Nature* **356,** 519–520 (1992).
22. Charlesworth, B., Morgan, M. T. & Charlesworth, D. The effect of deleterious mutations on neutral molecular variation. *Genetics* **134,** 1289–1303 (1993).
23. McDonald, J. H. & Kreitman, M. Adaptive protein evolution at the *Adh* locus in *Drosophila*. *Nature* **351,** 652–654 (1991).
24. Jenkins, D. L., Ortori, C. A. & Brookfield, J. F. A test for adaptive change in DNA sequences controlling transcription. *Proc. R. Soc. Lond. B* **261,** 203–207 (1995).
25. Egea, R., Casillas, S. & Barbadilla, A. Standard and generalized McDonald–Kreitman test: a website to detect selection by comparing different classes of DNA sites. *Nucleic Acids Res.* **36,** W157–W162 (2008).
26. Sawyer, S. A. & Hartl, D. L. Population genetics of polymorphism and divergence. *Genetics* **132,** 1161–1176 (1992).
27. Nielsen, R. Statistical tests of selective neutrality in the age of genomics. *Heredity* **86,** 641–647 (2001).
28. Eyre-Walker, A. Changing effective population size and the McDonald-Kreitman test. *Genetics* **162,** 2017–2024 (2002).
29. Charlesworth, J. & Eyre-Walker, A. The McDonald-Kreitman test and slightly deleterious mutations. *Mol. Biol. Evol.* **25,** 1007–1015 (2008).
30. Eyre-Walker, A. & Keightley, P. D. Estimating the rate of adaptive molecular evolution in the presence of slightly deleterious mutations and population size change. *Mol. Biol. Evol.* **26,** 2097–2108 (2009).
31. Fay, J. C., Wyckoff, G. J. & Wu, C. I. Testing the neutral theory of molecular evolution with genomic data from *Drosophila*. *Nature* **415,** 1024–1026 (2002).
32. Ohta, T. Slightly deleterious mutant substitutions in evolution. *Nature* **246,** 96–98 (1973).
33. David, J. R. & Capy, P. Genetic variation of *Drosophila melanogaster* natural populations. *Trends Genet.* **4,** 106–111 (1988).
34. Begun, D. J. & Aquadro, C. F. African and North American populations of *Drosophila melanogaster* are very different at the DNA level. *Nature* **365,** 548–550 (1993).
35. Tajima, F. Statistical methods to test for nucleotide mutation hypothesis by DNA polymorphism. *Genetics* **123,** 585–595 (1989).
36. Smith, N. G. & Eyre-Walker, A. Adaptive protein evolution in *Drosophila*. *Nature* **415,** 1022–1024 (2002).
37. Stoletzki, N. & Eyre-Walker, A. Estimation of the neutrality index. *Mol. Biol. Evol.* **28,** 63–70 (2011).
38. Ayroles, J. F. *et al.* Systems genetics of complex traits in *Drosophila melanogaster*. *Nature Genet.* **41,** 299–307 (2009).
39. Yamamoto, A. *et al.* Neurogenetic networks for startle-induced locomotion in *Drosophila melanogaster*. *Proc. Natl Acad. Sci. USA* **105,** 12393–12398 (2008).
40. Harbison, S. T., Yamamoto, A. H., Fanara, J. J., Norga, K. K. & Mackay, T. F. C. Quantitative trait loci affecting starvation resistance in *Drosophila melanogaster*. *Genetics* **166,** 1807–1823 (2004).
41. Pospisilik, J. A. *et al. Drosophila* genome-wide obesity screen reveals hedgehog as a determinant of brown versus white adipose cell fate. *Cell* **140,** 148–160 (2010).
42. Bellen, H. J. *et al.* The BDGP gene disruption project: single transposon insertions associated with 40% of *Drosophila* genes. *Genetics* **167,** 761–781 (2004).
43. The ModENCODE Consortium. Identification of functional elements and regulatory circuits by *Drosophila* modENCODE. *Science* **330,** 1787–1797 (2010).
44. Charlesworth, B., Coyne, J. A. & Barton, N. H. The relative rates of evolution of sex chromosomes and autosomes. *Am. Nat.* **130,** 113–146 (1987).
45. Sieberts, S. K. & Schadt, E. E. Moving toward a system genetics view of disease. *Mamm. Genome* **18,** 389–401 (2007).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements This work was supported by National Institutes of Health grant GM 45146 to T.F.C.M., E.A.S. and R.R.H.A.; R01 GM 059469 to R.R.H.A., MCI BFU 2009-09504 to A.B., R01 GM 085183 to K.R.T., NHGRI U54 HG003273 to R.A.G.; and an award through the NVIDIA Foundation's "Compute the Cure" programme to D.M.

Author Contributions T.F.C.M., S.R. and R.A.G. conceived the project. T.F.C.M., S.R., A.B. and E.A.S. wrote the main manuscript. T.F.C.M., S.R., A.B., E.A.S., J.F.A., K.R.T., J.M.C., C.M.B. and D.M. wrote the Supplementary methods. M.M.M., C.B., K.P.B., M.A.C., L.C., L.D., Y.H., M.J., J.C.J., S.N.J., K.W.J., F. Lara, F. Lawrence, S.L.L., R.F.L., M.M., D.M.M., L.N., I.M., L.P., L.L.P., C.Q., J.G.R., S.M.R., L.T., K.C.W., Y.-Q.W., A.Y. and Y.Z. performed experiments. T.F.C.M., A.B., J.F.A., D.Z., S.C., M.M.M., J.M.C., M.F.R., M.B., D.C., R.S.L., A.M., C.M.B., K.R.T., D.M. and E.A.S. did the bioinformatics and data analysis. J.F.A., S.C., M.M.M., Z.H., P.L., M.R., J.R. and E.A.S. wrote the Methods and did the web site development. R.R.H.A. contributed resources.

Author Information Sequences have been deposited at the National Center for Biotechnology Information Short Read Archives (http://www.ncbi.nlm.nih.gov/sra?term=DGRP). Reprints and permissions information is available at www.nature.com/reprints. This paper is distributed under the terms of the Creative Commons Attribution-Non-Commercial-Share Alike licence, and is freely available to all readers at www.nature.com/nature. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to T.F.C.M. (trudy_mackay@ncsu.edu).

B

Mycobacterial phylogenomics

Mycobacterial phylogenomics: An enhanced method for gene turnover analysis reveals uneven levels of gene gain and loss among species and gene families
Genome Biology and Evolution. Submitted

Mycobacterial phylogenomics: An enhanced method for gene turnover analysis reveals uneven levels of gene gain and loss among species and gene families

Pablo Librado[1, †], Filipe G. Vieira[1, 2, †], Alejandro Sánchez-Gracia[1, †] Sergios-Orestis Kolokotronis[3, 4, †], and Julio Rozas[1,*]

[1] Departament de Genètica & Institut de Recerca de la Biodiversitat (IRBio), Universitat de Barcelona, Barcelona, Spain

[2] Department of Integrative Biology, University of California, Berkeley, CA

[3] Department of Biological Sciences, Fordham University, Bronx, NY

[4] Sackler Institute for Comparative Genomics, American Museum of Natural History, New York, NY

*Corresponding author: E-mail: jrozas@ub.edu.

[†]These authors contributed equally to this work

Abstract

Species of the genus *Mycobacterium* differ in several features, from geographic ranges, and degree of pathogenicity, to ecological and host preferences. The recent availability of several fully sequenced genomes for a number of these species enabled the comparative study of the genetic determinants of this wide lifestyle diversity. Here, we applied two complementary phylogenetic-based approaches using information from 19 *Mycobacterium* genomes to obtain a more comprehensive view the evolution of this genus. First, we inferred the phylogenetic relationships using two new approaches, one based on amino acid substitutions but using a specific matrix for *Mycobacterium*, and the other based on a gene content dissimilarity matrix. We then, utilized our recently developed gain-and-death stochastic models to study gene turnover dynamics in this genus in a maximum likelihood framework. We uncovered a scenario that differs markedly from traditional 16S rRNA data and improves upon recent phylogenomic approaches. We also found that the rates of gene gain and death are high and unevenly distributed both across species and gene families, further supporting the utility of the new models of rate heterogeneity applied in a phylogenetic context. Finally, our Gene Ontology overrepresentation analysis among the most expanded or contracted gene families revealed that transposable elements and fatty acid metabolism-related gene families are likely the most important drivers of *Mycobacterium* genus diversification.

Keywords: Gene turnover rates, Gene gain-and-loss, Gene families, Maximum Likelihood, Rate heterogeneity, *M. tuberculosis*

Introduction

The genus Mycobacterium represents a large group of approx. 120-170 ecologically diverse type strains (American Type Culture Collection, http://www.atcc.org; Leibniz Institute DSMZ-German Collection of Microorganisms and Cell Cultures, http://www.dsmz.de; StrainInfo, http://www.straininfo.net/taxa/1059) that include both animal-adapted and free-living taxa. Although most of these species are ubiquitous environmental saprophytes, around one-third are pathogenic to vertebrates (Howard and Byrd 2000). These include members of the *M. tuberculosis* complex (MTBC), like *Mycobacterium tuberculosis* (the etiologic agent of tuberculosis and the major cause of death in AIDS patients worldwide) (Raviglione, et al. 1995), *M. bovis* (causes tuberculosis in cattle and other mammals) (O'Reilly and Daborn 1995; Une and Mori 2007) or *M. africanum* (infects humans specifically in West Africa) (de Jong, et al. 2010), as well as *M. leprae* (the causative agent of leprosy) (Sasaki, et al. 2001), *M. ulcerans* (responsible for the Buruli ulcer) (van der Werf, et al. 1999) or the *M. avium* complex (MAC), consisting of *M. avium* (opportunistic human pathogen) and *M. avium* ssp. *paratuberculosis* (causes Johne's disease in ruminants; see Johne's Information Center, University of Wisconsin, at http://www.johnesdisease.org and http://www.johnes.org), among others.

In recent years, there has been an emergence of tools and techniques for the rapid characterization of complete microbial genomes (Comas, et al. 2010; MacLean, et al. 2009). At present, we are able to exploit the powerful analytical methods of molecular evolution and population genomics to determine the relative contribution of the different

evolutionary forces that shape mycobacterial genome organization, structure and diversity. These methods offer the exceptional opportunity to explore the genetic and genomic determinants of pathogenesis, virulence, or lifestyle diversity in bacteria. Moreover, such analyses not only contribute to a better understanding of the biology of these species but also improve the diagnosis of mycobacterial diseases and the development of new strategies to identify potential drug targets and vaccine candidates.

The mutational events underlying gene and genome evolution are varied, ranging from nucleotide substitutions, to gene gains and losses, or changes in genomic structure and organization. Comparative genomic studies in the Mycobacterium genus have uncovered a genomic evolution characterized by frequent and unevenly distributed gene gain and loss events, being some of them associated with pathogenesis and virulence. Nevertheless, these studies applied either simple pairwise comparisons of genomic features (Gomez-Valero, et al. 2007), or used the gene tree and species tree reconciliation approach under a parsimony context (McGuire, et al. 2012). Therefore, the use of full probabilistic approaches that explicitly take heterogeneity into account (both across gene families and lineages) can, not only provide robust parameter estimates (gene turnover rates, ancestral gene content, number of gene gains and loss events, *etc.*), but also provide a rigorous statistical framework to contrast different biologically realistic scenarios. Nevertheless, the performance of maximum likelihood (ML) approaches in a phylogenetic context is highly dependent on the accuracy of the underlying species tree. When the tree is not simultaneously co-estimated, the analysis should be performed applying the most accurate topology and branch lengths. In recent years a number of

phylogenomic analyses have studied the phylogenetic relationships on the genus *Mycobacterium* (McGuire, et al. 2012; Prasanna and Mehra 2013; Smith, et al. 2012; Vishnoi, et al. 2010). However, these results do not agree on a single evolutionary history for the diversification of the genus, showing differences both in topology and branch lengths.

Here, we have estimated the phylogenetic relationships of the genus using two new approaches. One based on a *Mycobacterium* specific amino acid substitution matrix inferred from the proteome alignment of 19 genomes, and the other based on gene content dissimilarity built using the mutually exclusive genes among genomes. Taking advantage of this, we performed an exhaustive study of the gain and death dynamics, with special focus on gene turnover rate heterogeneity both across lineages and among families. The results uncover an evolutionary scenario that differs, to some extent, from most previously published molecular systematics studies, although it is in agreement with those who advocate the position of *M. leprae* as a sister taxon to the MTBC. We also show that mycobacterial evolution has been dominated by a high gene gain-and-loss dynamic with strong heterogeneity across lineages and gene families, especially for those families involved in transposition and fatty acid biosynthesis.

Methods

Mycobacterial Genomes and Orthologous Gene Identification

We retrieved the genome sequences of several *Mycobacterium* species publicly available on the Integr8 database (Kersey, et al. 2005) (Supplementary Table S1, Supplementary Material online), including their protein sequences and corresponding pairwise orthologous relationships. We included at least two strains per species of *M. tuberculosis*, *M. bovis* and *M. leprae*, which resulted in a total of 19 strains examined in this study: *M. tuberculosis* H37Rv; *M. tuberculosis* H37Ra; *M. tuberculosis* CDC1551; *M. bovis* BCG/Tokyo 172; *M. bovis* BCG/Pasteur 1173P2; *M. bovis* AF2122197; *M. leprae* TN; *M. leprae* Br4923; *M. marinum* M; *M. ulcerans* Agy99; *M. avium* 104; *M. paratuberculosis* K-10; *M.* sp. JLS; *M.* sp. KMS; *M.* sp. MCS; *M. vanbaaleni* PYR-1; *M. gilvum* PYR-GCK; *M. smegmatis* MC2 155; and *M. abscessus* ATCC 19977. To identify groups of orthologous sequences, we clustered the pairwise orthologous relationships using a Markov Clustering Algorithm (MCL) (Enright, et al. 2002) with an inflation value = 1.50.

Estimation of Substitution Matrices

The multiple sequence alignment (MSA) for each group of 1:1 orthologous protein sequences (one single gene copy per species; dataset Myc19, Supplementary Table S1) was built with MAFFT 6.603b using the L-INSi algorithm *(--localpair --maxiterate 1000*) (Katoh, et al. 2005). The MSA of the corresponding coding sequence (CDS) was performed according to a previous study (Vieira, et al. 2007) by back-translating the amino acid alignment using in-house developed Perl scripts.

The general time-reversible (GTR or REV) amino acid substitution matrix (Lanave, et al. 1984) of the 19-proteome alignment of 1:1 orthologs was independently estimated using a ML approach so that the time-reversibility condition $f_i q_{ij} = f_j q_{ji}$ was satisfied, where f is the amino acid equilibrium frequency and q is the exchangeability rate. A tree topology was estimated in the fine-grained parallel POSIX-threads build of RAxML 7-8 (Stamatakis 2006; Stamatakis and Ott 2008) by first using a stepwise-addition maximum parsimony starting tree; subsequently, a ML subtree-pruning-regrafting tree search was performed with the WAG amino acid substitution matrix (Whelan and Goldman 2001) (10 searches). This tree and the WAG matrix were used as a fixed topology and an initial set of substitution rates, respectively, for GTR matrix optimization in the *codeml* program of the PAML 4.3 package (Yang 2007).

Phylogenetic Analysis

The phylogenetic relationships among mycobacterial proteomes were estimated using the ML approach in RAxML, the newly generated GTR matrix (hereafter referred to as MYC for this specific amino acid dataset) and by accounting for among-site substitution rate heterogeneity by means of a discrete gamma distribution with four rate categories (Γ_4) (Yang 1994). Node support was evaluated with 500 bootstrap pseudo-replicates (Felsenstein 1985). To assess the fit of static replacement matrices commonly used for bacterial genomes, we performed the phylogenetic reconstruction in RAxML using the WAG, JTT (Jones, et al. 1992), and LG (Le and Gascuel 2008), including the Γ_4 model with observed amino acid frequencies (+F). The impact of substitution model choice in phylogenetic inference was quantified by comparing their log-likelihood. A partitioned

phylogenetic analysis was also carried out in RAxML by using the same substitution matrix across partitions while unlinking the amino acid frequencies, rate heterogeneity (different Γ distribution α shape parameter), and branch lengths (constrained to be proportional to the final best ML tree). We also inferred the ML phylogenetic tree of the 19 strains from 16S rRNA sequences in RAxML, mined from the genome assemblies of the examined taxa and aligned with MAFFT using the L-INSi algorithm.

In order to explore the putative topological discordance among trees of individual loci, we built ML trees for every set of 1:1 protein orthologs and subsequently constructed a consensus network (Holland, et al. 2004) in SplitsTree 4.13.1 (Huson and Bryant 2006) with edge weights reflecting the number of trees containing that edge and a threshold of 10%, meaning that the splits used to build that consensus network existed in at least 10% of the individual trees. We used our estimated MYC amino acid substitution matrix along with empirical amino acid frequencies calculated from every protein alignment. Individual orthologous protein tree support (PTS) was quantified by filtering the phylogenomic ML tree through the swarm of 1,011 individual protein trees. Tree-to-tree distances were estimated using the Robinson-Foulds metric (Robinson and Foulds 1981), as implemented in RAxML. Random unrooted trees (n = 5000) for 19 taxa were simulated in T-REX (Boc, et al. 2012).

Gene Content Dissimilarity

We used gene content dissimilarity to generate a distance-based phylogenetic tree. The proportion of mutually exclusive genes between taxa *a* and *b* is given by

$$d_{ab} = \frac{E_{ab}}{T_{ab}},$$

where E_{ab} is the total number of genes exclusive to taxa a and b, and T_{ab} is the total number of genes of these taxa. Operationally, we compute E_{ab} as $E_{ab} = T_a + T_b - 2S_{ab}$ and T_{ab} as $T_{ab} = T_a + T_b - S_{ab}$, where T_a, T_b, and S_{ab} are the total number of genes in taxon a, in taxon b, and the number of shared genes between these taxa, respectively. We clustered taxa based on their d_{ab} pairwise distances using a modified neighbor-joining algorithm, BioNJ (Gascuel 1997). Bootstrap analysis was performed by randomly sampling (with replacement) the gene families from the original data set. Node support values were calculated using SumTrees (Sukumaran and Holder 2010) from 500 bootstrap replicates.

Gene Family Evolution

We estimated the gene turnover rates with the ML method implemented in BadiRate v1.5 (Librado, et al. 2012). We used the Gain-and-Death (GD) stochastic model to estimate gain (γ) and death (δ) rates; this model is specifically suitable for the analysis of gene family evolution with high gene turnover rates or putative horizontal gene transfer (HGT). The analysis was conditioned on the rooted ultrametric tree (by using the *M. abscessus* lineage as ancestor of all other *Mycobacterium* species), derived from our estimated phylogenomic ML tree (Figure 1). We fitted four different branch models to our data (Myc18 dataset, Supplementary Table S1, Supplementary Material online): (i) a global rates (GD-GR-ML) model, where both the gain and death turnover rates are constant over time; (ii) a free-rates (GD-FR-ML) model that assumes separate GD rates for each branch; (iii) a pathogen-specific rates model (GD-PR-ML), which allows for two

branch classes (one for pathogenic and another for non-pathogenic lineages); and, finally, (iv) the PR model (GD-PRl-ML) in which the branch leading to the *M. leprae* clade has its own GD rates. In addition, we also fitted a free rates (FR) model that takes into account GD heterogeneity across gene families using two discrete Γ distributions with two categories (GD -FR-ML+dG2). The goodness of fit of these models was assessed using likelihood-ratio tests (LRT) and the Akaike Information Criterion (AIC) (Akaike 1974). We used the best-fit branch model to estimate the ancestral gene content of each family and the γ and δ rates. We examined the function of the families with significantly high or low turnover rates using gene ontology (GO) term enrichment analysis as implemented in Ontologizer 2.0.

Results

Amino acid-based Phylogenetic Tree

The ML phylogenetic tree of the 19 mycobacteria was estimated using information of 1,011 1:1 orthologs (364,491 amino acids; Figure 1) and our estimated MYC-GTR amino acid substitution matrix (Supplementary Figure S1, Supplementary Material online). Overall, it is in concordance with the major relationships known for this genus (Tortoli 2003): MTBC is sister to *M. leprae* with close affinities to *M. marinum* and *M. ulcerans*, and the two *M. avium* strains follow in a ladderized tree order. A sister clade to all the above comprises the rapid-growing, non-pathogenic mycobacteria *M. vanbaalenii*, *M. gilvum* and *M. smegmatis*, as well as the free-living environmental strains.

These relationships are markedly different from those obtained on the 16S rRNA gene phylogeny (Supplementary Figure S2, Supplementary Material online). Although 16S rRNA supports the monophyly of slow and rapid growers at 82% and 100%, it provides weak to moderate bootstrap support for mid-depth nodes, like the monophyletic *M. leprae*+*M. avium* clade, and the MTBC+(*M. leprae*+*M. avium*) relationship (<66%). Similarly, low levels of support are evident for the placement of *M. smegmatis*, *M. vanbaaleni*, and *M. gilvum* (52-65%). Copy number variation for this gene has been reported at both the intraspecific (Acinas, et al. 2004) and interspecific (Vetrovsky and Baldrian 2013) levels. Here, we found slow growers to harbor a single 16S rRNA gene copy, while rapid growers contained two copies that were physically separated by well over 1 Mb of their assembled genomes. Copies were mostly identical within species with the exception of *M. vanbaalenii* and the free-living soil strains JLS, KMS, and MCS

(Supplementary Figure S2, Supplementary Material online). However, using one vs. both 16S rRNA gene copies did not impact phylogenetic inference or node stability, with the sole exception of the free-living strains that formed paraphyletic assemblages within their clade, due to extreme sequence conservation in that lineage.

Because the amino acid substitution matrix can also impact the phylogenetic inference (Le and Gascuel 2008), we estimated the specific amino acid replacement matrix for the 19 *Mycobacterium* proteome dataset (MYC matrix; Supplementary Figure S1, Supplementary Material online). The specific MYC matrix has important differences with respect to the WAG matrix, both in amino acid equilibrium frequencies and replacement rates. Indeed, the MYC matrix has a higher frequency of Ala, Arg, Cys, and Val and a lower frequency of Asn, Ile, Lys, Phe, Ser, and Tyr. Most exchangeability rates are highly correlated between MYC and WAG, and most differences were owed to radical changes that were more frequent in MYC. Yet, the trees that were built using MYC, WAG, LG and JTT exhibit identical topologies, although the MYC-based tree has a better fit to the data (likelihood scores: -2,947,171 for MYC vs. -2,960,115 for LG+F, -2,966,562 for JTT+F, -2,962,299 for WAG+F). Partitioning the alignment into 1,011 loci yielded an even better likelihood score of -2,937,575.

Internal edges on the consensus network of 1,011 protein ML trees showed reticulation, thus indicating disagreement among splits (Supplementary Figure S2B, Supplementary Material online). Although the overall phylogenetic structure of the network is similar to that of the phylogenomic ML tree with terminal nodes being generally supported by the vast majority of protein trees, there seems to be localized incongruence (Supplementary

Figure S2A, Supplementary Material online). Specifically, for up to a 25% threshold of splits reticulation this incongruence is still apparent, thus providing uncertainty on the bifurcating relationships among the MTBC and *M. leprae* (PTS = 264), while 366 protein trees support an *M. marinum*+*M. ulcerans*–MTBC relationship, and 253 trees favor an *M. leprae*–*M. avium* relationship. Similarly, a reticulation remains among the free-living strain group, *M. gilvum*+*M. vanbaalenii*, and *M. smegmatis* with the two alternative groupings being equally supported by orthologs (PTS = 310 for MCS-KMS-JLS–*M. gilvum*+*M. vanbaalenii*, and PTS = 296 for *M. smegmatis*–MCS-KMS-JLS). Once our threshold became more stringent, i.e. 30%, the reticulations disappeared and a bifurcating topology emerged. Interestingly, this topology differs from the phylogenomic one, as 366 trees favor an MTBC–*M. marinum*+*M. ulcerans* grouping. This indicates that the individual trees causing topological conflict express the history of a minority. Overall, bipartition conflict was variable with only 12 out of 510,555 total pairwise protein tree comparisons exhibiting no disagreement, and, on the other end of the conflict spectrum, 997 comparisons with trees differing at every split.

These tree support levels are also noticeable when we contrasted the phylogenomic tree to all gene trees (Supplementary Figure S2A, Supplementary Material online), however this method of tree agreement quantification does not report the alternative splits that are in conflict with the reference – in this case the phylogenomic topology. In order to harvest information on the alternative splits, e.g. second and third most prevalent, consensus network approaches are needed. In spite of the topological conflict among protein trees and between the phylogenomic tree and protein trees, the orthologs examined here appear to have related genealogical histories overall, as indicated by the

low average relative Robinson-Foulds (RF) distance of 0.46 (i.e. any two protein trees differed by 14.80 ± 3.25 out of 32 splits on average) compared to the RF distance of 0.989 calculated from 5000 simulated random 19-taxon trees.

Gene content dissimilarity-based tree

The evolutionary dynamics of gene gain and loss and amino acid replacements in orthologous genes can reflect different aspects of genome evolution, especially in the case of mycobacteria. The number of identifiable 1:1 orthologs when considering the 19 genomes is very low (an effect enlarged by the inclusion of a genome with a radically different number of genes, as *M. leprae*), revealing the importance of gene turnover in these species. Then, the information based on the amino acid replacements among orthologs may provide only a partial view of the evolution of the genus. To explore the effect of gene content on the phylogenetic relationships among mycobacteria, we devised a method to capture gene content dissimilarity as a distance measure (see Methods). Interestingly, gene content-based analysis (Figure 2), which makes use of a different type of genetic information, recapitulates the same topology that the ML phylogenomic protein sequence analysis based on our newly estimated MYC amino acid replacement matrix, except for the MTBC strains. In fact, the analysis showed an increased resolution between close relationships, especially among the strains of this complex. In particular, the H37Rv genome exhibited the highest gene content dissimilarity with respect to the other members of the clade, and the analysis uncovered a ladder-like array of relationships that correlates with the total number of genes as well as corroborates the emergence of *M. bovis* out of *M. tuberculosis* (with a 98% of bootstrap support for this node) (Smith, et al. 2006). *M. bovis* strains are in fact more closely related to the virulent

H37Rv than to other *M. tuberculosis* strains (100% of bootstrap support). As expected the gene content-based tree clearly reflects the massive gene loss in the two *M. leprae* strains, which are a very special case (with only 1,603 and 1,599 genes). Moreover, the comparison of gene number between *M. tuberculosis* (3,949) and *M. leprae* (1,605) genes along with their respective genome sizes (4.40 and 3.27 Mb, respectively) indicates a marked discrepancy between genome and gene content dynamics (χ^2 test; $P = 2.20 \times 10^{-16}$), showing that mycobacterial taxa with a similar total number of genes may harbor strikingly different gene repertoires.

Gene Gain and Loss Dynamics in the Evolution of Mycobacteria

To gain insights into the major evolutionary mechanisms that act on mycobacterial evolution, we analyzed the distribution of gene gains and losses within a phylogenetic context using the statistical framework provided by BadiRate (Librado, et al. 2012). We fixed the tree topology to the inferred under our newly estimated MYC matrix and estimated the GD rates, as well as, the number of genes in each internal phylogenetic node. Strikingly, the GD turnover rates were not only extremely high but also unevenly distributed among lineages (after excluding unreliable estimates of short branches, the gain rates ranged from 0 to 1.07 gene gains/substitution, with death rates from 6.00×10^{-3} to 4.01 losses/gene/substitution; the highest death rate corresponding to *M. leprae*). In fact, the statistical comparison among the four branch-based assessments showed that GD-FR-ML is the model that best fits the observed gene content data (Table 1; Supplementary Figure S3, Supplementary Material online). The mere separation between pathogenic and non-pathogenic evolutionary dynamics (GD-PR-ML model) does not

sufficiently explain the high GD rates exhibited by these genomes, even after accounting for the singularities of the *M. leprae* lineage (GD-PRl-ML model). The branches of the MTBC clade also show very different GD rates relative to both *M. leprae* lineages and to the rest of the pathogenic group branches.

To understand the biological meaning of such high gene turnover rates, we analyzed the putative heterogeneity of GD rates across gene families. We calculated the likelihood of the data under a model that takes into account rate heterogeneity both across lineages and gene families (GD-FR-ML+dG2), which fitted the observed data significantly better than the FR model (ΔAIC = 2,158.13). Interestingly, the estimates of the shape parameters of the discrete Gamma distribution used to model rate heterogeneity among families (α_{gain} = 0.08, α_{death} = 0.99) indicate that gain and death processes are also heterogeneously distributed (Figure 4).

Combining phylogenomic and gene gain-and-loss information, we estimated that the most recent common ancestor (MRCA) of the MTB clade likely had an intermediate number of genes (3,053) than previously predicted (2,977) (Gomez-Valero, et al. 2007) (3,901) (McGuire, et al. 2012). Our analysis demonstrates that despite sharing a very similar number of genes (especially for closely related species and subspecific strains), the species-specific gene repertoire is very different; for instance, the genomes of *M. ulcerans* and *M. tuberculosis*, harboring 4,206 and 3,990 genes, respectively, only share approximately half (48.30%) of their genes. Results also show a global trend of gene number increases across the diversification of the *Mycobacterium* taxa (Figure 3). Indeed,

when excluding both *M. leprae* and short (likely unreliable) branches, no lineage exhibits a clear net reduction in gene content. Remarkably, the non-pathogenic *Mycobacterium* lineages appear to have gained many more genes than pathogenic and this effect is not explained by large expansions in specific branches, but rather equally observed in both internal and external branches of this clade.

To gain insight into the functional meaning of such high gene turnover rates, we analyzed the overrepresented GO categories among outlier families (i.e., families with rates that depart significantly from the estimated GD process). Given the best-fit model (GD-FR-ML+dG2), we found outliers in 15 of the 34 lineages (Supplementary Table S2, Supplementary Material online), in both internal and external branches. Notably, most of the outlier gene families (33) displayed gain rates that were significantly higher than expected (on a given phylogenetic branch), whereas only one was significantly contracted. Specific GO categories (for biological processes) were found to be statistically overrepresented in 6 of the 34 analyzed branches (Figure 3), with transposition/DNA recombination and fatty acid biosynthesis being the most frequent GO term (found in 2 lineages each). This result points to transposable elements and fatty acid metabolism as the most important determinants of high gene turnover rates in mycobacteria.

Interestingly, some of the above mentioned outlier gene families with overrepresented GO terms have been previously related to virulence in these bacteria (Figure 3) (Forrellad, et al. 2013). In particular, gene families that encode important integral

membrane components of the cell wall (NanT) (Ioerger, et al. 2013), polyketide synthases (Pks) (Li, et al. 2010) and enzymes responsible for the assembly of mycobactin (Mbt) (Quadri, et al. 1998) displayed higher than expected gain rates in three of the pathogenic lineages. Among outlier families without significantly overrepresented GO terms (or with overrepresented terms of molecular function or cellular component), some *PE/PPE*- and *mmpL*-related genes (Domenech, et al. 2005; Mukhopadhyay and Balaji 2011) exhibited higher gain rates in *M. marinum*. On the other hand, a family of genes related to fatty acid hydrolases (*Mmcs1433*), some endonucleases (genes related to *Mvan0273*), haloacetate dehalogenases (genes related to *Msmeg1984* and *Msmeg2340*) and hydrolases (genes related to *Mmar2844*) were outliers of gain rates in some lineages of the non-pathogenic clade.

Discussion

Our comparative genomic analysis of mycobacteria improves upon the knowledge provided by previous studies that have focused on this group. First, we have estimated a specific amino acid replacement matrix (MYC) and obtained a more robust phylogeny, which is critical to the study of gene content and gene family turnover rates in a phylogenetic context. Second, we applied, for the first time, stochastic models of gene gain and death to the study of mycobacteria genome evolution under a ML framework; this approach provides a sound probabilistic framework to contrast different evolutionary scenarios and to separately estimate gene gain and loss rates as well as ancestral gene content.

Phylogenetic Relationships

Our phylogenomic framework not only allows for the proposal of a robust evolutionary scenario for the diversification of mycobacteria, but also provides a trustworthy tree for subsequent phylogenetic analyses. In our ML tree, the different strains of *M. tuberculosis* and *M. bovis* form a single, well-defined clade, and pathogenic and non-pathogenic lineages are clearly separated. Nevertheless, both under a partitioned and an unpartitioned scheme, we uncovered some important topological differences with respect to most of the relationships published for this genus (Smith, et al. 2012), specially the position of non-MTBC species (*M. marinum*, *M. ulcerans* and *M. avium*) and of some non-pathogenic species. These topological differences seem to be data-driven, since our 16S rRNA ML tree (Supplementary Figure S2, Supplementary Material online) recovered a different topology. This locus has been considered as a “gold standard” for bacterial molecular

systematics (Janda and Abbott 2007) since the days of DNA-DNA hybridization, but in spite of its functional and sequence conservation, it is now apparent that in the case of *Mycobacterium* as a genus, it does not reflect genome-level evolution. Four main topological differences emerged from the comparison of our ML tree and the 16S rRNA tree: (1) the separation of the *leprae-avium* clade (sister to MTBC) and the placement of the *avium* clade basal to all other slow growers; (2) the transfer of the *marinum-ulcerans* clade from a sister relationship with *M. leprae* to a more basal position outside the *M. leprae*-MTBC clade; (3) *M. smegmatis* was repositioned from sister taxon of the free-living soil strains to the base of the rapid growers; (4) *M. gilvum* and *M. vanbaalenii* were pulled together in a well-supported clade sister to the soil strains. Our ML model-based approach also differs from Vishnoi, et al. (2010) in the placement of *M. leprae* and the non-MTBC pathogenic species and from McGuire, et al. (2012) in the phylogenetic relationships between *M. smegmatis* and the rest of non-pathogenic mycobacteria. The choice of amino acid substitution model did not impact topology inference, but our estimated MYC GTR model yielded a better likelihood score. Interestingly, our phylogenomic tree was not fully supported by the individual orthologs. Not unlike other studies spanning a variety of organisms from fungi to vertebrates (Gatesy and Baker 2005; Salichos and Rokas 2013), we found orthologs exhibiting a conflicting phylogenetic history when compared to one another as well as to the concatenated proteome phylogenomic tree. While shallow nodes are supported by the vast majority of orthologs, deeper alternative groupings emerge when individual locus trees are contrasted. Concatenation produced a fully bootstrap-supported phylogeny at the interspecific level, while masking individual tree-to-tree disagreement that was made

evident by adopting a network approach that enabled us to examine alternate groupings on a reticulated topology. Reasons for such incongruity include stochasticity, e.g. incomplete lineage sorting, varying sequence convergence levels, lack of phylogenetic informativeness, recombination, as well as horizontal gene transfer. By combining concatenation with locus-specific phylogenetic methods we were able to propose a robust total-evidence evolutionary scenario, while dissecting contradictory evolutionary signals at the gene level.

Gene Content-based analysis

Although they are highly conserved at the sequence level throughout the genus, the number identifiable 1:1 orthologs in *Mycobacterium* taxa is very low. This pattern reflects the major impact of gene content changes, rather than amino acid substitutions, in genome evolution. This is especially true for the MTBC, where our gene-content phylogeny provides much higher resolution than the amino acid substitution matrix-based approach. Here, our analysis clearly supports (with high bootstrap support values) the origin of the *M. bovis* strains from a *M. tuberculosis*-like ancestor, being the H37Rv genome as the closest relative in terms of gene content. The topology for the rest of the genus is identical to the ML tree, confirming the inferences from our phylogenomic analysis and supporting our topology in contrast to the inferred in other studies (McGuire, et al. 2012; Vishnoi, et al. 2010).

Gain-and-Death Dynamics

To study in more detail the gene dynamics across the genus, we applied a full likelihood method to study *Mycobacterium* genome evolution. The applied models not only allow for the decoupled estimation of gene gain and death rates (i.e., as two independently estimated parameters, γ and δ) but also explicitly take into account key features of the evolution of *Mycobacterium* species, such as horizontal gene transfer (i.e., gain rates that include new copies that originated from 0 ancestral genes) and, more importantly, the heterogeneity of GD rates across lineages and gene families. The use of these ML-based models allows for the statistical evaluation of competing evolutionary scenarios and the selection of the one that best explains observed gene dynamics.

In this study, we demonstrate that *Mycobacterium* species show high gene turnover rates that differ markedly across lineages and families (GD-FR-ML+dG2 is the best fit model) and not merely between pathogenic and non-pathogenic lineages, like several previous studies assumed in their analyses (McGuire, et al. 2012; Prasanna and Mehra 2013). Clearly, ignoring these rate heterogeneities can greatly bias estimates and the identification of particular outlier lineages and families, especially when using parsimony-based approaches. We found that our ML estimates of the number of genes at the ancestral nodes clearly differ with respect to that found by a previous study (McGuire, et al. 2012), which were based on a gene trees species tree reconciliation method implemented in the SYNERGY algorithm (Wapinski, et al. 2007). These strong differences can likely be explained by the use of a more realistic (and complex) model under a solid statistical framework. We cannot rule out the possibility, however, that the previously reported low phylogenetic coverage of non-mycobacteria *Actinobacteria* (McGuire, et al. 2012) could also have a significant influence on their results.

The analysis of gene families with extreme GD rates may also be sensitive to the model assumed. In fact, the number of genes of a particular can differ between species simply by the stochastic turnover process, i.e., without significant rate changes. Therefore, the mere comparison of family sizes is not sufficient to detect important family expansions or contractions. Our methodology, however, allows for the detection of these outliers within the estimated gain-and-death stochastic background. In fact, we have been able to identify specific lineages in which the gain or loss of gene family members might have significant biological relevance. Noticeably, we found manifest differences between pathogenic and non-pathogenic lineages, with the latter presenting outlier families related to pathogenesis in *M. tuberculosis*. It has been documented that variations in the proteins and lipids that form the cell wall are major determinants of virulence in the MTBC (Forrellad, et al. 2013; Reddy, et al. 2013; Sonawane, et al. 2012). These virulence factors serve as targets for antimicrobial drug development (North, et al. 2013; Tomioka, et al. 2011; Wang, et al. 2013). Here, we found that the most important gene gains in fatty acid and siderophore biosynthesis (Campbell and Cronan 2001; Rodriguez 2006) and carbohydrate membrane transport (Niederweis 2008; Titgemeyer, et al. 2007) likely took place in the ancestral lineage that led to the pathogenic clade, after the split of *M. marinum* and *M. ulcerans*. Despite their close phylogenetic relationship, the two branches connecting *M. marinum* and *M. ulcerans* had accumulated gene families with unexpectedly high gene dynamism. Some of these gene families likely contributed to the important phenotypic differences that are observed between these two species (Stinear, et al. 2008; Yip, et al. 2007). The most dynamic gene families in non-pathogenic strains are enzymes (e.g. metal-dependent hydrolases, haloacetate dehalogenases, endonucleases,

isopentenyl pyrophosphate isomerases) and transcriptional regulator encoding families. These results suggest that the diversification and adaptation of non-pathogenic mycobacteria to their different life-styles may have been promoted by changes in some of these gene families.

Literature Cited

Acinas SG, Marcelino LA, Klepac-Ceraj V, Polz MF. 2004. Divergence and redundancy of 16S rRNA sequences in genomes with multiple rrn operons. J Bacteriol 186: 2629-2635.

Akaike H. 1974. A new look at the statistical identification model. {IEEE} Trans Automat Contr 19: 716-723.

Boc A, Diallo AB, Makarenkov V. 2012. T-REX: a web server for inferring, validating and visualizing phylogenetic trees and networks. Nucleic Acids Res 40(W1):W573-W579.

Campbell JW, Cronan JE, Jr. 2001. Bacterial fatty acid biosynthesis: targets for antibacterial drug discovery. Annu Rev Microbiol 55: 305-332. doi: 10.1146/annurev.micro.55.1.305

Comas I, et al. 2010. Human T cell epitopes of *Mycobacterium tuberculosis* are evolutionarily hyperconserved. Nature Genet 42: 498-503. doi: 10.1038/ng.590

de Jong BC, Antonio M, Gagneux S. 2010. *Mycobacterium africanum*--review of an important cause of human tuberculosis in West Africa. PLoS Neglect Trop Dis 4: e744. doi: 10.1371/journal.pntd.0000744

Domenech P, Reed MB, Barry CE. 2005. Contribution of the *Mycobacterium tuberculosis* MmpL protein family to virulence and drug resistance. Infect Immun 73: 3492-3501. doi: 10.1128/IAI.73.6.3492-3501.2005

Enright AJ, Van Dongen S, Ouzounis CA. 2002. An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res 30: 1575-1584.

Felsenstein J. 1985. Confidence limits on phylogenies: an approach using the bootstrap.

Evolution 39: 783-791.

Forrellad MA, et al. 2013. Virulence factors of the *Mycobacterium tuberculosis* complex. Virulence 4: 3-66. doi: 10.4161/viru.22329

Gascuel O. 1997. BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. Mol Biol Evol 14: 685-695.

Gatesy J, Baker RH. 2005. Hidden likelihood support in genomic data: can forty-five wrongs make a right? Syst Biol 54:483-492.

Gomez-Valero L, Rocha EP, Latorre A, Silva FJ. 2007. Reconstructing the ancestor of *Mycobacterium leprae*: the dynamics of gene loss and genome reduction. Genome Res 17: 1178-1185. doi: 10.1101/gr.6360207

Holland B, Huber KT, Moulton V, Lockhart P. 2004. Using consensus networks to visualize contradictory evidence for species phylogeny. Mol Biol Evol 21:1459-1461.

Howard ST, Byrd TF. 2000. The rapidly growing mycobacteria: saprophytes and parasites. Microbes Infect 2: 1845-1853.

Huson DH, Bryant D. 2006. Application of phylogenetic networks in evolutionary studies. Mol Biol Evol 23:254-267.

Ioerger TR, et al. 2013. Identification of new drug targets and resistance mechanisms in *Mycobacterium tuberculosis*. PloS One 8: e75245. doi: 10.1371/journal.pone.0075245

Janda JM, Abbott SL. 2007. 16S rRNA gene sequencing for bacterial identification in the diagnostic laboratory: pluses, perils, and pitfalls. J Clin Microbiol 45: 2761-2764. doi: 10.1128/JCM.01228-07

Jones DT, Taylor WR, Thornton JM. 1992. The rapid generation of mutation data matrices from protein sequences. Computer applications in the biosciences : CABIOS 8:

275-282.

Katoh K, Kuma K, Toh H, Miyata T. 2005. MAFFT version 5: improvement in accuracy of multiple sequence alignment. Nucleic Acids Res 33: 511-518. doi: 10.1093/nar/gki198

Kersey P, et al. 2005. Integr8 and Genome Reviews: integrated views of complete genomes and proteomes. Nucleic Acids Res 33: D297-302. doi: 10.1093/nar/gki039

Lanave C, Preparata G, Saccone C, Serio G. 1984. A new method for calculating evolutionary substitution rates. J Mol Evol 20: 86-93.

Le SQ, Gascuel O. 2008. An improved general amino acid replacement matrix. Mol Biol Evol 25: 1307-1320. doi: 10.1093/molbev/msn067

Li YJ, Danelishvili L, Wagner D, Petrofsky M, Bermudez LE. 2010. Identification of virulence determinants of *Mycobacterium avium* that impact on the ability to resist host killing mechanisms. J Med Microbiol 59: 8-16. doi: 10.1099/jmm.0.012864-0

Librado P, Vieira FG, Rozas J. 2012. BadiRate: estimating family turnover rates by likelihood-based methods. Bioinformatics 28: 279-281. doi: 10.1093/bioinformatics/btr623

MacLean D, Jones JD, Studholme DJ. 2009. Application of 'next-generation' sequencing technologies to microbial genetics. Nature Rev Microbiol 7: 287-296. doi: 10.1038/nrmicro2122

McGuire AM, et al. 2012. Comparative analysis of Mycobacterium and related Actinomycetes yields insight into the evolution of *Mycobacterium tuberculosis* pathogenesis. BMC Genomics 13: 120. doi: 10.1186/1471-2164-13-120

Mukhopadhyay S, Balaji KN. 2011. The PE and PPE proteins of *Mycobacterium tuberculosis*. Tuberculosis 91: 441-447. doi: 10.1016/j.tube.2011.04.004

Niederweis M. 2008. Nutrient acquisition by mycobacteria. Microbiology 154: 679-692. doi: 10.1099/mic.0.2007/012872-0

North EJ, Jackson M, Lee RE. 2013. New approaches to target the mycolic acid biosynthesis pathway for the development of tuberculosis therapeutics. Curr Pharm Desin. DOI: 10.2174/13816128196661311182036 41

O'Reilly LM, Daborn CJ. 1995. The epidemiology of *Mycobacterium bovis* infections in animals and man: a review. Tubercle and Lung Disease 76 Suppl 1: 1-46.

Prasanna AN, Mehra S. 2013. Comparative phylogenomics of pathogenic and non-pathogenic mycobacterium. PloS One 8: e71248. doi: 10.1371/journal.pone.0071248

Quadri LE, Sello J, Keating TA, Weinreb PH, Walsh CT. 1998. Identification of a *Mycobacterium tuberculosis* gene cluster encoding the biosynthetic enzymes for assembly of the virulence-conferring siderophore mycobactin. Chem Biol 5: 631-645.

Raviglione MC, Snider DE, Jr., Kochi A. 1995. Global epidemiology of tuberculosis. Morbidity and mortality of a worldwide epidemic. JAMA 273: 220-226.

Reddy PV, et al. 2013. Disruption of mycobactin biosynthesis leads to attenuation of *Mycobacterium tuberculosis* for growth and virulence. J Infect Dis 208: 1255-1265. doi: 10.1093/infdis/jit250

Robinson DR, Foulds LR. 1981. Comparison of phylogenetic trees. Math Biosci 53:131-147.

Rodriguez GM. 2006. Control of iron metabolism in *Mycobacterium tuberculosis*. Trends Microbiol 14: 320-327. doi: 10.1016/j.tim.2006.05.006

Salichos L, Rokas A. 2013. Inferring ancient divergences requires genes with strong phylogenetic signals. Nature 497:327-331.

Sasaki S, Takeshita F, Okuda K, Ishii N. 2001. *Mycobacterium leprae* and leprosy: a compendium. Microbiol Immunol 45: 729-736.

Smith NH, Gordon SV, de la Rua-Domenech R, Clifton-Hadley RS, Hewinson RG. 2006. Bottlenecks and broomsticks: the molecular evolution of *Mycobacterium bovis*. Nature Rev Microbiol 4: 670-681. doi: 10.1038/nrmicro1472

Smith SE, et al. 2012. Comparative genomic and phylogenetic approaches to characterize the role of genetic recombination in mycobacterial evolution. PloS One 7: e50070. doi: 10.1371/journal.pone.0050070

Sonawane A, Mohanty S, Jagannathan L, Bekolay A, Banerjee S. 2012. Role of glycans and glycoproteins in disease development by *Mycobacterium tuberculosis.* Crit Rev Microbiol 38: 250-266. doi: 10.3109/1040841X.2011.653550

Stamatakis A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22: 2688-2690. doi: 10.1093/bioinformatics/btl446

Stamatakis A, Ott M. 2008. Efficient computation of the phylogenetic likelihood function on multi-gene alignments and multi-core architectures. Philos Trans R Soc Lond B Biol Sci 363: 3977-3984. doi: 10.1098/rstb.2008.0163

Stinear TP, et al. 2008. Insights from the complete genome sequence of *Mycobacterium marinum* on the evolution of *Mycobacterium tuberculosis*. Genome Res18: 729 - 741.

Sukumaran J, Holder MT. 2010. DendroPy: a Python library for phylogenetic computing. Bioinformatics 26: 1569-1571. doi: 10.1093/bioinformatics/btq228

Titgemeyer F, et al. 2007. A genomic view of sugar transport in *Mycobacterium smegmatis* and *Mycobacterium tuberculosis*. J Bacteriol 189: 5903-5915. doi:

10.1128/JB.00257-07

Tomioka H, Tatano Y, Sano C, Shimizu T. 2011. Development of new antituberculous drugs based on bacterial virulence factors interfering with host cytokine networks. J Infect Chemother 17: 302-317. doi: 10.1007/s10156-010-0177-y

Tortoli E. 2003. Impact of genotypic studies on mycobacterial taxonomy: the new mycobacteria of the 1990s. Clin Microbiol Rev 16: 319-354.

Une Y, Mori T. 2007. Tuberculosis as a zoonosis from a veterinary perspective. Comp Immunol Microbiol Infect Dis 30: 415-425. doi: 10.1016/j.cimid.2007.05.002

van der Werf TS, van der Graaf WT, Tappero JW, Asiedu K. 1999. *Mycobacterium ulcerans* infection. Lancet 354: 1013-1018. doi: 10.1016/S0140-6736(99)01156-3

Vetrovsky T, Baldrian P. 2013. The variability of the 16S rRNA gene in bacterial genomes and its consequences for bacterial community analyses. PloS One 8: e57923. doi: 10.1371/journal.pone.0057923

Vieira FG, Sanchez-Gracia A, Rozas J. 2007. Comparative genomic analysis of the odorant-binding protein family in 12 Drosophila genomes: purifying selection and birth-and-death evolution. Genome Biol 8: R235. doi: 10.1186/gb-2007-8-11-r235

Vishnoi A, Roy R, Prasad HK, Bhattacharya A. 2010. Anchor-based whole genome phylogeny (ABWGP): a tool for inferring evolutionary relationship among closely related microorganisms [corrected]. PloS One 5: e14159. doi: 10.1371/journal.pone.0014159

Wang F, et al. 2013. Identification of a small molecule with activity against drug-resistant and persistent tuberculosis. Proc Natl Acad Sci U S A 110: E2510-2517. doi: 10.1073/pnas.1309171110

Wapinski I, Pfeffer A, Friedman N, Regev A. 2007. Automatic genome-wide

reconstruction of phylogenetic gene trees. Bioinformatics 23: i549-558. doi: 10.1093/bioinformatics/btm193

Whelan S, 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.

Yang Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. J Mol Evol 39: 306-314.

Yang Z 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol 24: 1586-1591. doi: 10.1093/molbev/msm088

Yip MJ, et al. 2007. Evolution of *Mycobacterium ulcerans* and other mycolactone-producing mycobacteria from a common *Mycobacterium marinum* progenitor. J Bacteriol 189: 2021-2029. doi: 10.1128/JB.01442-06

Table list

TABLE 1. Results of the branch models of gene turnover fitted to the gene families present in the genus *Mycobacterium*.

Figure Legends

FIG. 1. Maximum likelihood phylogenetic reconstruction of the relationships among the 19 *Mycobacterium* taxa inferred from 1011 1:1 orthologous protein sequences. Phylogenomic tree based on the MYC substitution matrix estimated from the concatenation of 1,011 mycobacterial orthologs. Each ortholog partition was allowed to evolve under a different Γ_4 model of among-site rate heterogeneity and its individual observed amino acid frequencies. All nodes received 100% bootstrap support. Red and green bars indicate pathogenic and non-pathogenic mycobacteria, respectively, while yellow and blue bars denote slow and rapid growth species, respectively.

FIG. 2. Gene content differences among 19 mycobacterial genomes. The upper diagonal matrix contains the number of genes shared between pairs of genomes, while the lower diagonal shows gene content similarities. The neighbor joining phylogram on the left clusters mycobacterial taxa by gene content dissimilarity. All nodes received 100% bootstrap support except for those denoted on the tree.

FIG. 3. Gain and death events under the best-fit model (GD-FR-ML+dG2). Numbers in internal nodes indicate the number of ancestral genes. Numbers on the branches denote the minimum gain (blue) and loss (red) events estimated from the data. Green, yellow, orange, and purple branches indicate the lineages where the outlier families (GD rates) are enriched in the biological processes of carbohydrate transmembrane transport, transposition, fatty acid biosynthesis, and the siderophore biosynthetic process,

respectively.

FIG. 4. Gain (blue) and death (red) rates of each lineage and each family category under the GD-FR-ML+dG2 model.

Supplementary Tables

TABLE S1. Datasets used for the analyses.

TABLE S2. Lineage-specific distribution of outlier families (GD rates) after fitting the GD-FR-ML+dG2 model (one representative of each family).

Supplementary Figures

FIG. S1. Properties of the GTR MYC amino acid substitution matrix and comparison with other matrices. (A) Bubble plot of the exchangeability rate differences between the WAG (blue) and MYC (red) amino acid substitution matrices. (B) Comparison of amino acid frequencies among different matrices (MYC, WAG, LG, JTT).

FIG. S2. Phylogenetic conflict among the concatenated proteome, 16S rRNA gene, and 1011 orthologous protein trees. (A) Topological differences between the phylogenetic reconstruction based on proteomes and the commonly used bacterial marker locus 16S rRNA. Values at the nodes indicate the bootstrap support values expressed as the proportion (%) of bootstrap trees that agree with a given bipartition on the best ML tree. Values in dark boxes indicate the number of orthologous protein trees that agree with a given bipartition on the best ML tree. *M. abscessus*, which can cause sporadic lung disease, was used as the outgroup. (B) Consensus network built with the 1,011 individual ML amino acid-based trees with a 10% threshold. Scale bar expresses the number of trees containing a given split.

FIG. S3. Number of genes estimated in ancestral nodes under the four branch-specific gain and death rate models. GD-GR-ML, global rates model; GD-PR-ML, pathogenic lineage-specific rates model; GD-PRl-ML, GD-PR-ML model + *M. leprae* specific rates model; GD-FR-ML, free rates model.

Table 1. Results of the branch models of gene turnover fitted to the gene families present in the genus *Mycobacterium*.

Branch model	No. of parameters	Log-likelihood score	AIC score	ΔAIC
GD-GR-ML	3	-90,718.03	181,442.06	29,783.16
GD-PR-ML	5	-89,185.81	178,381.62	26,722.72
GD-PRl-ML	7	-87,408.80	174,831.60	23,172.70
GD-FR-ML	69	-75,756.45	151,650.90	0.00

GD-GR-ML, Global GD rates model

GD-PR-ML, Pathogenic lineage-specific GD rates model

GD-PRl-ML, GD-PR-ML model + *M. leprae* lineage-specific GD rates

GD-FR-ML, Free GD rates model

Figure 1

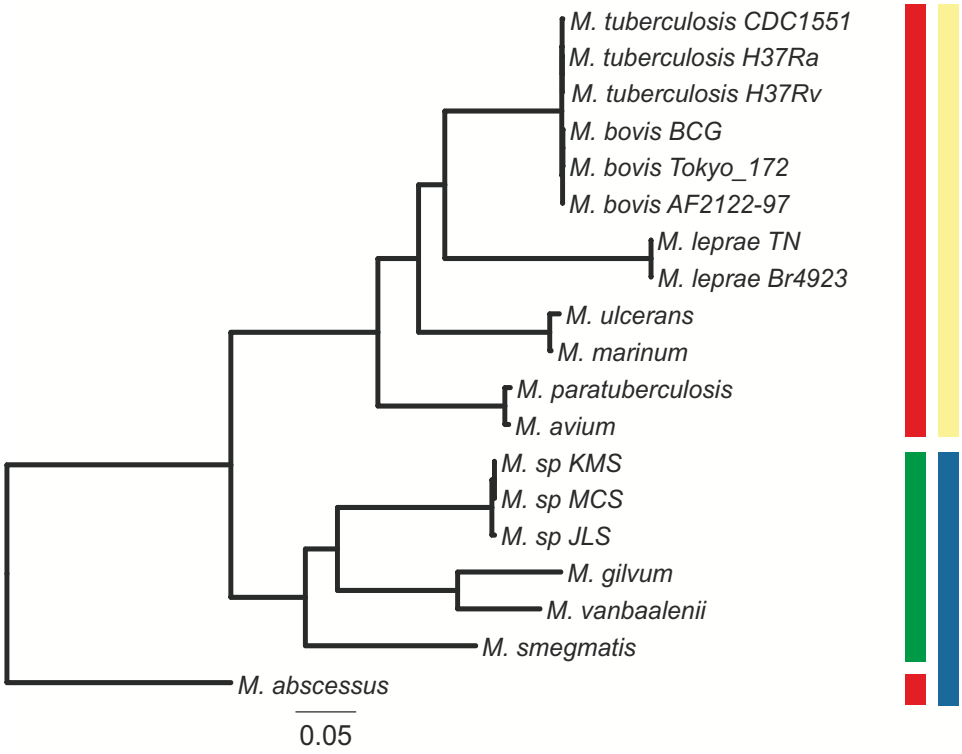

M. tuberculosis CDC1551
M. tuberculosis H37Ra
M. tuberculosis H37Rv
M. bovis BCG
M. bovis Tokyo_172
M. bovis AF2122-97
M. leprae TN
M. leprae Br4923
M. ulcerans
M. marinum
M. paratuberculosis
M. avium
M. sp KMS
M. sp MCS
M. sp JLS
M. gilvum
M. vanbaalenii
M. smegmatis
M. abscessus
0.05

Figure 2

M. tuberculosis CDC1551 (4198)
M. tuberculosis H37Ra (3990)
M. tuberculosis H37Rv (3949)
M. bovis Tokyo_172 (3906)
M. bovis BCG (3891)
M. bovis AF2122-97 (3911)
M. leprae TN (1603)
M. leprae Br4923 (1599)
M. marinum (5418)
M. ulcerans (4206)
M. paratuberculosis (4316)
M. avium (5040)
M. sp MCS (5588)
M. sp KMS (5893)
M. sp JLS (5695)
M. vanbaalenii (5499)
M. gilvum (5902)
M. smegmatis (6597)
M.abscessus (4939)
99
98
99

Figure 3

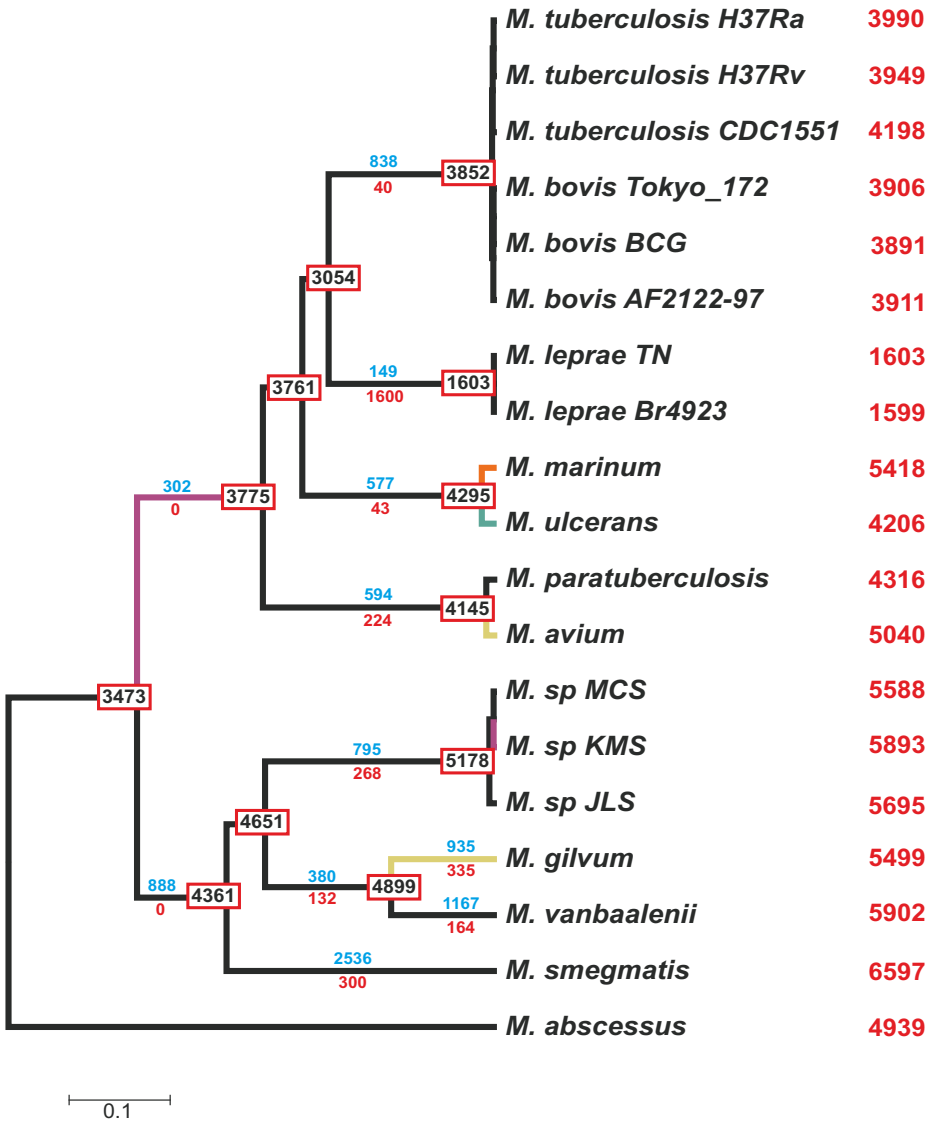
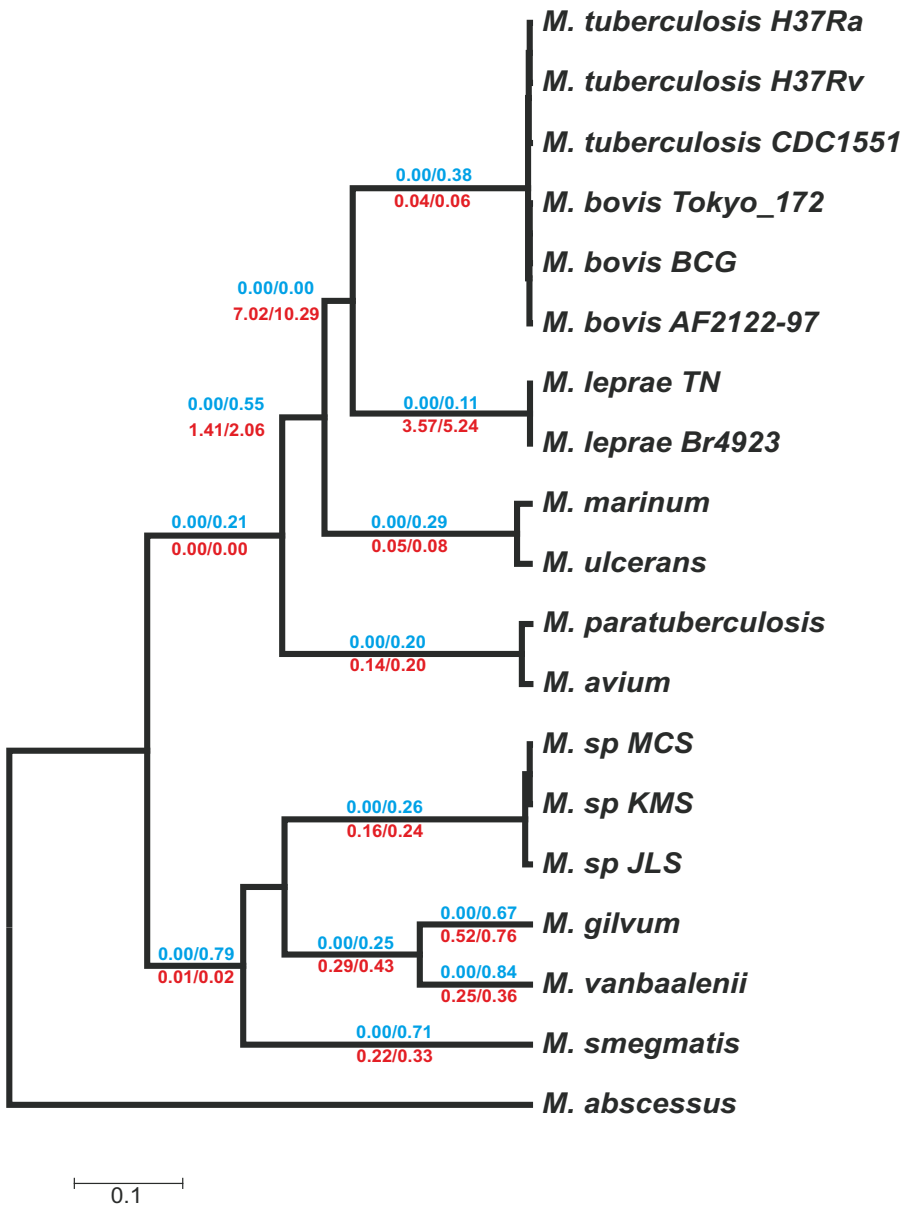

M. tuberculosis H37Ra
3990
M. tuberculosis H37Rv
3949
M. tuberculosis CDC1551
4198
838
40
3852
M. bovis Tokyo_172
3906
M. bovis BCG
3891
3054
M. bovis AF2122-97
3911
M. leprae TN
1603
149
1600
1603
3761
M. leprae Br4923
1599
M. marinum
5418
302
0
3775
577
43
4295
M. ulcerans
4206
M. paratuberculosis
4316
594
224
4145
M. avium
5040
3473
M. sp MCS
5588
M. sp KMS
5893
795
268
5178
M. sp JLS
5695
4651
935
335
M. gilvum
5499
888
0
4361
380
132
4899
1167
164
M. vanbaalenii
5902
2536
300
M. smegmatis
6597
M. abscessus
4939
0.1

Figure 4


M. tuberculosis H37Ra
M. tuberculosis H37Rv
M. tuberculosis CDC1551
M. bovis Tokyo_172
M. bovis BCG
M. bovis AF2122-97
M. leprae TN
M. leprae Br4923
M. marinum
M. ulcerans
M. paratuberculosis
M. avium
M. sp MCS
M. sp KMS
M. sp JLS
M. gilvum
M. vanbaalenii
M. smegmatis
M. abscessus
0.00/0.38
0.04/0.06
0.00/0.00
7.02/10.29
0.00/0.55
1.41/2.06
0.00/0.11
3.57/5.24
0.00/0.21
0.00/0.00
0.00/0.29
0.05/0.08
0.00/0.20
0.14/0.20
0.00/0.26
0.16/0.24
0.00/0.67
0.52/0.76
0.00/0.79
0.01/0.02
0.00/0.25
0.29/0.43
0.00/0.84
0.25/0.36
0.00/0.71
0.22/0.33
0.1

Supplementary Material

Mycobacterial phylogenomics: An enhanced method analysis for gene turnover analysis reveals uneven levels of gene gain and loss among species and gene families

Pablo Librado1, †, Filipe G. Vieira1, 2, †, Alejandro Sánchez-Gracia1, † Sergios-Orestis Kolokotronis3, 4, †, and Julio Rozas1

1 Departament de Genètica & Institut de Recerca de la Biodiversitat (IRBio), Universitat de Barcelona, Barcelona, Spain
2 Department of Integrative Biology, University of California, Berkeley, CA
3 Department of Biological Sciences, Fordham University, Bronx, NY
4 Sackler Institute for Comparative Genomics, American Museum of Natural History, New York, NY

Table S1. Data sets used for the analyses

Data set	Analysis	Genomes	Markers and data information
Myc19	Phylogenetic reconstruction	The 19 mycobacteria proteomes (*n* = 19) Data Set	Concatenation of 1,011 aligned protein sequences (1:1 orthologs; 364,491 aminoacid positions) Data Set MYC amino acid substitution matrix
Myc18	Gain and death analysis	The Myc19 data set, excluding *M. abscessus* (*n* = 18)	14,108 ortholog groups (N:N and N:M relationships)

Table S2

lineage*	DataBase	ID	Symbol	GOID	DB:reference	Evidence	coWith/From	Aspect	ObjectName	ObjectSynonym	ObjectType	Taxon	Date	AssignedBy
10_8	UniProtKB	A5U6L0												
18_16	UniProtKB	B2HEJ6												
18_16	UniProtKB/TrEMBL	B2HFL0	MMAR_1396	GO:0003677	GO_REF:0000002	IEA	InterPro:IPR002559	F	TransposaseforinsertionMMAR_1396|MMAR_14		protein	taxon:216594	20090716	UniProtKB
18_16	UniProtKB	B2HIY0												
18_16	UniProtKB/Swiss-Prot	O07798	pks2	GO:0008152	GO_REF:0000002	IEA	InterPro:IPR016038	P	Phthioceranic/hydroxyphpks2|msl-2|Rv3825c|MT		protein	taxon:1773	20090716	UniProtKB
18_16	UniProtKB	O53268												
18_16	UniProtKB	O86329	mbtE	GO:0000036	GO_REF:0000002	IEA	InterPro:IPR009081	F	PEPTIDESYNTHETASEMBO86329_MYCTU|Rv2380		protein	taxon:1773	20111015	InterPro
18_16	UniProtKB/Swiss-Prot	P95211	mmpL1	GO:0016020	GO_REF:0000002	IEA	InterPro:IPR004869	C	Putativemembraneprot mmpL1|Rv0402c|MTCY0		protein	taxon:1773	20090716	UniProtKB
18_17	UniProtKB	O07730	nanT	GO:0005215	GO_REF:0000002	IEA	InterPro:IPR005829	F	PROBABLESIALICACID-TO07730_MYCTU|nanT|R		protein	taxon:1773	20111015	InterPro
22_21	UniProtKB	O06367	Rv3640c	GO:0003677	GO_REF:0000002	IEA	InterPro:IPR001207	F	PROBABLETRANSPOSASO06367_MYCTU|Rv3640		protein	taxon:1773	20111015	InterPro
22_21	UniProtKB/Swiss-Prot	P0C5G8	Rv0796	GO:0032196	GO_REF:0000004	IEA	SP_KW:KW-0815	P	PutativetransposaseforMT0413|MT3100|Rv079		protein	taxon:1773	20090722	UniProtKB
23_19	UniProtKB/Swiss-Prot	O07798	pks2	GO:0008152	GO_REF:0000002	IEA	InterPro:IPR016038	P	Phthioceranic/hydroxyphpks2|msl-2|Rv3825c|MT		protein	taxon:1773	20090716	UniProtKB
23_19	UniProtKB	P95243	MT2412	GO:0005576	PMID:16973880	IDA		C	PutativeESAT-6-likeprotES6L7_MYCTU|MTCY98.		protein	taxon:1773	20100414	MTBBASE
23_19	UniProtKB/Swiss-Prot	P95246	plcB	GO:0016787	GO_REF:0000004	IEA	SP_KW:KW-0378	F	PhospholipaseC2 plcB|mpcB|Rv2350c|MT		protein	taxon:1773	20090722	UniProtKB
23_19	UniProtKB/Swiss-Prot	Q10550	Rv0890c	GO:0003677	GO_REF:0000004	IEA	SP_KW:KW-0238	F	PutativeHTH-typetranscRv0890c|MTCY31.18c|M		protein	taxon:1773	20090722	UniProtKB
23_19	UniProtKB	Q6MX44												
23_22	UniProtKB	P64885												
26_25	UniProtKB/TrEMBL	Q1BC39	Mmcs_1433	GO:0006633	GO_REF:0000002	IEA	InterPro:IPR006694	P	PutativeuncharacterizedMmcs_1433		protein	taxon:164756	20090716	UniProtKB
28_27	UniProtKB/TrEMBL	A3Q477	Mjls_4182	GO:0006350	GO_REF:0000004	IEA	SP_KW:KW-0804	P	TranscriptionalregulatorMjls_4182		protein	taxon:164757	20090722	UniProtKB
3_1	UniProtKB	P95243	MT2412	GO:0005576	PMID:16973880	IDA		C	PutativeESAT-6-likeprotES6L7_MYCTU|MTCY98.		protein	taxon:1773	20100414	MTBBASE
31_29	UniProtKB	A4T3B4												
31_29	UniProtKB/TrEMBL	A4TDW6	Mflv_4283	GO:0004803	GO_REF:0000002	IEA	InterPro:IPR001207	F	Transposase,mutatortypMflv_4283		protein	taxon:350054	20090716	UniProtKB
31_29	UniProtKB/Swiss-Prot	Q10621	Rv1313c	GO:0003677	GO_REF:0000004	IEA	SP_KW:KW-0238	F	TransposaseforinsertionRv1313c|MTCY373.33c|		protein	taxon:1773	20090722	UniProtKB
31_30	UniProtKB/TrEMBL	A1T1S2	Mvan_0273	GO:0003676	GO_REF:0000002	IEA	InterPro:IPR002711	F	PutativeuncharacterizedMvan_0273		protein	taxon:350058	20090716	UniProtKB
31_30	UniProtKB	A1UPH2												
31_30	UniProtKB/TrEMBL	B2HE04	MMAR_2844	GO:0016810	GO_REF:0000002	IEA	InterPro:IPR011059	F	Metal-dependenthydrolMMAR_2844		protein	taxon:216594	20090716	UniProtKB
32_31	UniProtKB	P60230	Rv2512c	GO:0003677	GO_REF:0000002	IEA	InterPro:IPR001207	F	TransposaseforinsertionTRA1_MYCTU|MTCI364.		protein	taxon:1773	20111015	InterPro
32_31	UniProtKB	P63689	ctpG	GO:0000166	GO_REF:0000002	IEA	InterPro:IPR008250	F	Probablecation-transporCTPG_MYCTU|ctpG|Rv1		protein	taxon:1773	20111015	InterPro
34_32	UniProtKB	B2HGI8												
34_33	UniProtKB/TrEMBL	A0QTW0	MSMEG_1984	GO:0018785	GO_REF:0000003	IEA	EC:3.8.1.3	F	HaloacetatedehalogenasMSME G_1984		protein	taxon:246196	20090717	UniProtKB
34_33	UniProtKB/TrEMBL	A0QUU7	MSMEG_2340	GO:0005737	GO_REF:0000002	IEA	InterPro:IPR011179	C	PutativeuncharacterizedMSMEG_2340		protein	taxon:246196	20090716	UniProtKB
34_33	UniProtKB	B2HFN8												
34_33	UniProtKB/TrEMBL	P71977	MT1758	GO:0005622	GO_REF:0000002	IEA	InterPro:IPR001808	C	PROBABLETRANSCRIPTIOMT1758|Rv1719		protein	taxon:1773	20090716	UniProtKB
35_23	UniProtKB/Swiss-Prot	O07798	pks2	GO:0008152	GO_REF:0000002	IEA	InterPro:IPR016038	P	Phthioceranic/hydroxyphpks2|msl-2|Rv3825c|MT		protein	taxon:1773	20090716	UniProtKB

http://www.geneontology.org/GO.format.gaf-1_0.shtml

Tree*

(((((((29397_1,30_2)3,53_4)5,((33099_6,25994_7)8,138_9)10)11,(29_12,32549_13)14)15,(30990_16,25870_17)18)19,(679_20,25829_21)22)23,(((25058_24,25946_25)26,28710_27)28,(29236_29,25929_30)31)32,25827_33)34)35;

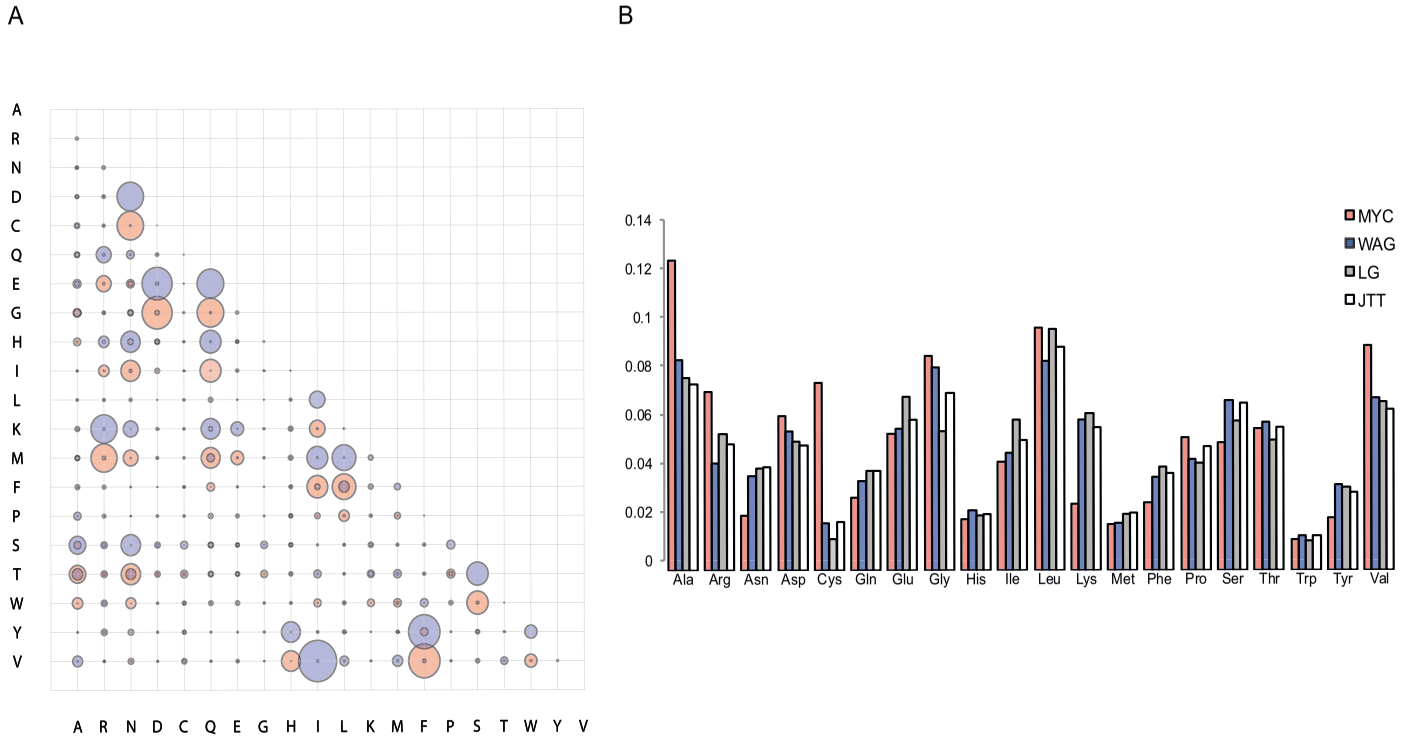

A
A R N D C Q E G H I L K M F P S T W Y V
B
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
MYC
WAG
LG
JTT
Ala Arg Asn Asp Cys Gln Glu Gly His Ile Leu Lys Met Phe Pro Ser Thr Trp Tyr Val

Figure S1

Figure S2A

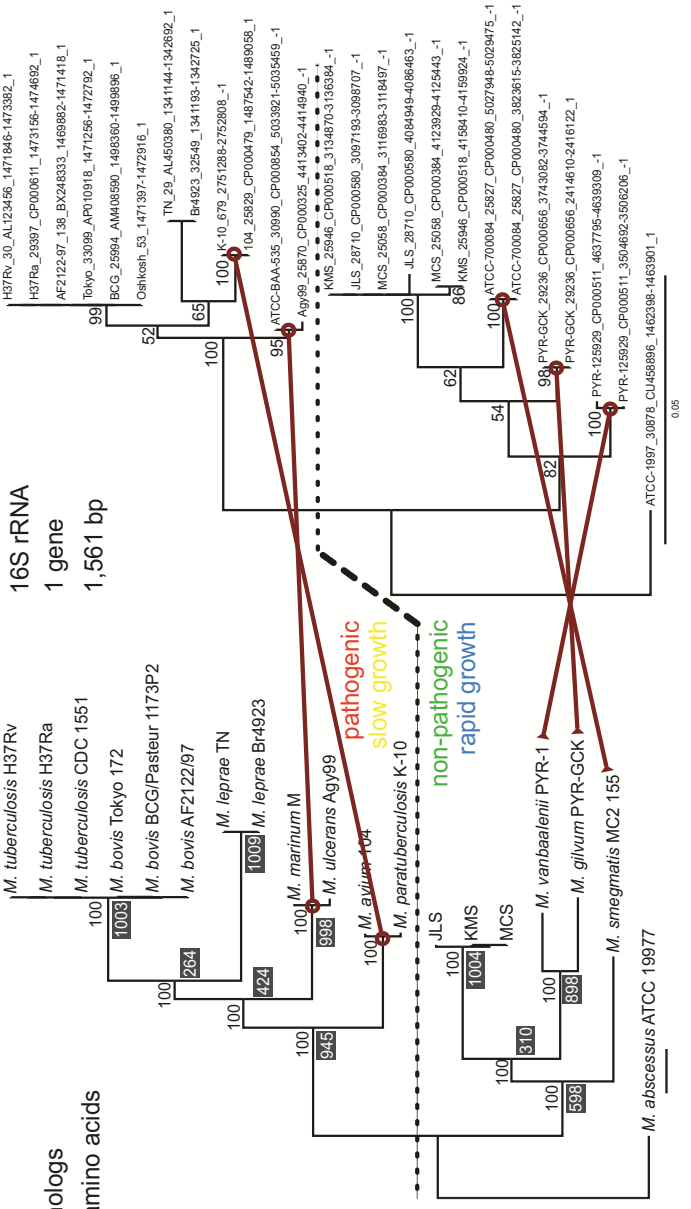
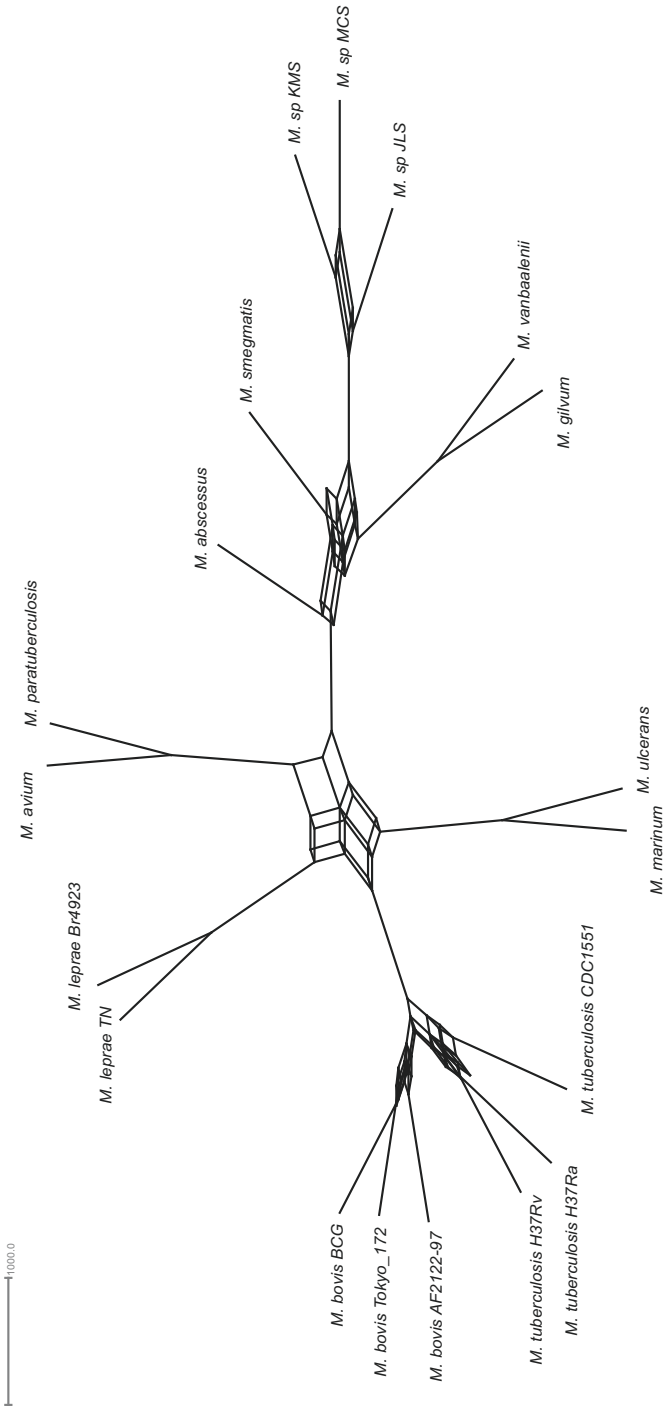

Genomes
1,011 orthologs
364,491 amino acids
16S rRNA
1 gene
1,561 bp
M. tuberculosis H37Rv
M. tuberculosis H37Ra
M. tuberculosis CDC 1551
M. bovis Tokyo 172
M. bovis BCG/Pasteur 1173P2
M. bovis AF2122/97
M. leprae TN
M. leprae Br4923
M. marinum M
M. ulcerans Agy99
M. avium 104
M. paratuberculosis K-10
JLS
KMS
MCS
M. vanbaalenii PYR-1
M. gilvum PYR-GCK
M. smegmatis MC2 155
M. abscessus ATCC 19977
pathogenic
slow growth
non-pathogenic
rapid growth
H37Rv_30_AL123456_1471846-1473382_1
H37Ra_29397_CP000611_1473156-1474692_1
AF2122-97_138_BX248333_1469882-1471418_1
Tokyo_33099_AP010918_1471256-1472792_1
BCG_25994_AM408590_1498360-1499896_1
Oshikosh_53_1471397-1472916_1
TN_29_AL450380_1341144-1342692_1
Br4923_32549_1341193-1342725_1
K-10_679_2751286-2752808_-1
104_25829_CP000479_1487542-1489058_1
ATCC-BAA-535_30990_CP000854_5033921-5035459_-1
Agy99_25870_CP000325_4413402-4414940_-1
KMS_25946_CP000518_3134870-3136384_-1
JLS_28710_CP000580_3097193-3098707_-1
MCS_25058_CP000384_3116983-3118497_-1
JLS_28710_CP000580_4084949-4086463_-1
MCS_25058_CP000384_4123929-4125443_-1
KMS_25946_CP000518_4158410-4159924_-1
ATCC-700084_25827_CP000480_5027948-5029475_-1
ATCC-700084_25827_CP000480_3823615-3825142_-1
PYR-GCK_29236_CP000656_3743082-3744594_-1
PYR-GCK_29236_CP000656_2414610-2416122_1
PYR-125929_CP000511_4637795-4639309_-1
PYR-125929_CP000511_3504692-3506206_-1
ATCC-1997_30878_CU458896_1462398-1463901_1
0.05

Figure S2B


M. sp KMS
M. sp MCS
M. sp JLS
M. smegmatis
M. vanbaalenii
M. gilvum
M. abscessus
M. paratuberculosis
M. avium
M. ulcerans
M. marinum
M. leprae Br4923
M. leprae TN
M. tuberculosis CDC1551
M. bovis BCG
M. bovis Tokyo_172
M. bovis AF2122-97
M. tuberculosis H37Rv
M. tuberculosis H37Ra
1000.0

C
Informe del director

Universitat de Barcelona

Dr. Julio Rozas
Catedràtic de Genètica

Departament de Genètica
Facultat de Biologia

Diagonal 643
Edifici Prevosti
08028 Barcelona

Tel. + 34 93 4021495
Fax. + 34 93 4034420
jrozas@ub.edu
www.ub.edu/molevol/julio

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

---

El Dr. **Julio Rozas Liras,** director de la Tesi Doctoral elaborada pel Sr. **Pablo Librado Sanz**, amb el títol "**Genómica evolutiva de la regulación transcripcional en las principales familias multigénicas del sistema quimiosensorial** de ***Drosophila***"

INFORMA

Que la tesi doctoral està elaborada com a compendi de 5 publicacions amb dades originals (publicacions 1-5 en el cos central de la tesi), i dos més (publicacions 6-7) a l'apèndix:

1. Librado, P. and Rozas, J. 2009. DnaSP v5: A software for comprehensive analysis of DNA polymorphism data. *Bioinformatics* **25:** 1451-1452.
   Factor d'impacte: **4.926**. Ocupa la posició **2** (sobre **29**) dins la categoria de Mathematical and Computational Biology.
2. Librado, P., Vieira, F. G. and Rozas, J. 2012. BadiRate: Estimating Family Turnover Rates by Likelihood-Based Methods. *Bioinformatics* **28:** 279-281.
   Factor d'impacte: **5.332**. Ocupa la posició **2** (sobre **47**) dins la categoria de Mathematical and Computational Biology
3. Ramia*, M., Librado*, P., Casillas*, S., Rozas, J. and Barbadilla, A. 2012. PopDrowser: the Population Drosophila Browser. *Bioinformatics* **28:** 595-596.
   *, la mateixa contribució
   Factor d'impacte: **5.332**. Ocupa la posició **2** (sobre **47**) dins la categoria de Mathematical and Computational Biology
4. Librado, P. and Rozas, J. 2013. Uncovering the Functional Constraints Underlying the Genomic Organization of the Odorant-Binding Protein Genes. *Genome Biol. Evol.* **5:** 2096-2108.
   Factor d'impacte: **4.759 (**any 2012**)**. Ocupa la posició **27** (sobre **161**) dins la categoria de Genetics and Heredity
5. Librado, P. and Rozas, J. 2014. Positive selection drives the evolution of the transcriptional regulatory upstream regions of the major chemosensory gene families.
   En preparació.

6. Mackay, T. F. C. et al. 2012 (including P. Librado and J. Rozas). The Drosophila melanogaster Genetic Reference Panel. *Nature* **482:** 173-178.
   Factor d'impacte: **38.597**. Ocupa la posició **1** (sobre **56**) dins la categoria de Multidisciplinary Sciences.

Dr. Julio Rozas
Catedràtic de Genètica

Departament de Genètica
Facultat de Biologia

Diagonal 643
Edifici Prevosti
08028 Barcelona

Tel. + 34 93 4021495
Fax. + 34 93 4034420
jrozas@ub.edu
www.ub.edu/molevol/julio

7. Librado*, P., Vieira*, F. G., Sánchez-Gracia*, A., Kolokotronis*, S. O. and Rozas, J. Mycobacterial phylogenomics: An enhanced method for gene turnover analysis reveals uneven levels of gene gain and loss among species and gene families.
   *, la mateixa contribució
   Enviat a publicar.

A les publicacions 1-2 i 4-5 el doctorand va realitzar la feina computacional i d'anàlisi de dades, i va redactar el primer esborrany dels manuscrits. A la publicació 3 (col·laboració amb el grup de A. Barbadilla; UAB) va col·laborar en el disseny del browser genòmic i a les tasques computacionals. A la publicació 6, on participen diversos grups de recerca, va participar en la implementació del browser genòmic. A la publicació 7 (col·laboració amb membres del grup de recerca de J. Rozas) va participar amb les feines computacionals i d'anàlisi de dades. En cap cas s'ha utilitzat, implícitament o explícitament, els treballs presentats en el cos central d'aquesta tesi per a l'elaboració d'una altra tesi doctoral.

Dr. Julio Rozas Liras

Catedràtic de Genètica

Universitat de Barcelona

D
Financiación

Esta tesis doctoral ha estado financiada por los proyectos BFU2007-62927 y BFU2010-15484 del Ministerio de Educación y Ciencia, y por los proyectos 2005SGR-00166 y 2009SGR-1287 de la Comissió Interdepartamental de Recerca i Innovació Tecnològica. Durante el período de formación pre-doctoral, Pablo Librado Sanz ha disfrutado de una Beca de Formación del Personal Investigador (FPI; BES-2008-2059) del Ministerio de Educación y Ciencia.