



UNIVERSITAT DE
BARCELONA

Evolución molecular de los genes del grupo Polycomb en el género *Drosophila*

Juan Manuel Calvo Martín



Aquesta tesi doctoral està subjecta a la llicència **Reconeixement- NoComercial – CompartirIgual 4.0. Espanya de Creative Commons.**

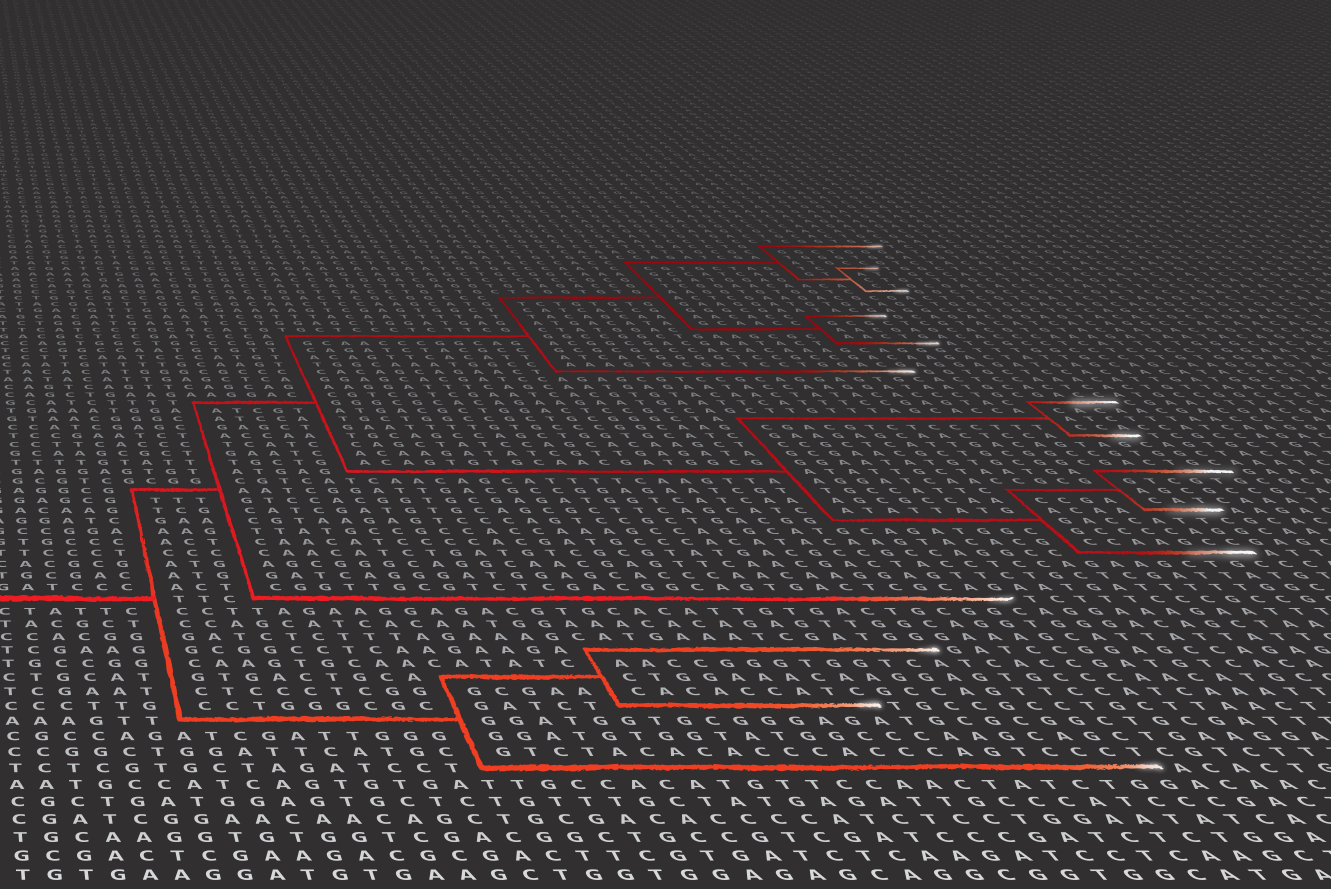
Esta tesis doctoral está sujeta a la licencia **Reconocimiento - NoComercial – CompartirIgual 4.0. España de Creative Commons.**

This doctoral thesis is licensed under the **Creative Commons Attribution-NonCommercial-ShareAlike 4.0. Spain License.**

Evolución molecular de los genes del grupo **Polycomb** en el género *Drosophila*

Juan Manuel Calvo Martín

Tesis doctoral 2017





UNIVERSITAT ^{DE}
BARCELONA

Evolución molecular de los genes del grupo Polycomb en el género *Drosophila*

Memoria presentada por **Juan Manuel Calvo Martín**
para optar al Grado de Doctor por la Universidad de Barcelona.

Departamento de Genética, Microbiología y Estadística

El autor de la tesis

Juan Manuel Calvo Martín

La tutora y codirectora de la tesis

Dra. Carmen Segarra

Profesora Titular
Departamento de Genética,
Microbiología y Estadística
Facultad de Biología
Universidad de Barcelona

La codirectora de la tesis

Dra. Montserrat Papaceit

Profesora Agregada
Departamento de Genética,
Microbiología y Estadística
Facultad de Biología
Universidad de Barcelona

Barcelona, Abril de 2017

“

... el té en casa de Teodoro se convirtió en costumbre semanal. Todos los jueves Spiro me llevaba al pueblo, con los bolsillos reventando de cajas de fósforos y tubos de ensayo llenos de ejemplares. Era una cita a la que no habría faltado por nada del mundo.

G. Durrell

Mi familia y otros animales

*A Carmen y Montse, por ser compañeras
además de jefas.*

*A Montse Aguadé y Julio Rozas, por liderar
un grupo en el que es un placer trabajar.*

*A David, por dejarme usar su laboratorio,
porque es suyo.*

*A Pablo, Rosser y Eva, por marcarme el
camino, a Cris por andarlo conmigo, y a la
muchachada que viene tras nosotros.*

*A mis padres, por absolutamente todo...
y a ti también, sí.*

*A los amigos, que entienden lo que hago lo
justo para estar someramente orgullosos.*

*Y a Nuria, que prácticamente me lo impuso
como requisito indispensable.*

Índice

1	Introducción
3	Evolución molecular
3	Breve historia de la evolución molecular
4	Teoría neutralista y evolución a diferentes escalas
6	Tras la huella de la selección
10	Distintas aproximaciones: del gen al genoma
12	Los genes del grupo Polycomb
12	Función de las proteínas Polycomb
14	Los complejos Polycomb
17	<i>Polycomb response elements</i>
19	<i>Drosophila</i> como organismo de estudio
19	El género <i>Drosophila</i>
20	El grupo <i>obscura</i> y <i>Drosophila subobscura</i>
22	Evolución de los genes Polycomb en <i>Drosophila</i> : consideraciones previas
	...

25	Objetivos
29	Informe de las directoras
33	Artículos
35	Artículo 1
37	Adaptive selection and coevolution at the proteins of the Polycomb repressive complexes in <i>Drosophila</i>
53	Artículo 2
55	Evidence of neofunctionalization after the duplication of the highly conserved Polycomb group gene <i>Caf1-55</i> in the obscura group of <i>Drosophila</i>
73	Artículo 3
75	Molecular population genetics of the Polycomb genes in <i>Drosophila subobscura</i>
115	Artículo 4
117	<i>Sex combs reduced</i> (<i>Scr</i>) regulatory region of <i>Drosophila</i> revisited
139	Discusión
142	Divergencia y coevolución de las proteínas Polycomb
144	Identificación de un nuevo parálogo: <i>Caf1-55dup</i>
146	Inferencia de selección adaptativa
149	Caracterización de la región reguladora del gen Hox <i>Scr</i>
151	Conclusiones
157	Bibliografía

Introducción

Evolución molecular

Breve historia de la evolución molecular

La evolución molecular es el proceso de cambio de las secuencias de ADN a lo largo de las generaciones, siendo los cambios introducidos en las poblaciones durante dicho proceso la causa última de la biodiversidad existente en nuestro planeta. Para entender el origen de esta disciplina científica debemos remontarnos a mediados del siglo XIX, cuando se establecieron los dos pilares sobre los que se sustenta el estudio de la evolución biológica. En primer lugar, A. R. Wallace y C. Darwin definieron el proceso evolutivo y su principal mecanismo de acción: la selección natural. Tal como se describe en la obra seminal *El origen de las especies* (Darwin 1859), la selección natural se basa en que aquellos individuos con características que les permitan adaptarse mejor al medio tendrán mayor probabilidad de sobrevivir y, por tanto, de transmitir dichas características a la descendencia. En segundo lugar y con apenas unos años de diferencia, G. Mendel estableció las leyes que rigen la herencia de estas características, fundando así la ciencia de la genética (Mendel 1866).

En las primeras décadas del siglo XX el estudio de la herencia mendeliana a nivel poblacional desencadenó el nacimiento de la genética de poblaciones, cuyo corpus teórico y matemático fue desarrollado por R. A. Fisher, S. Wright y J. B. S. Haldane, conciliando la genética mendeliana con la idea de la evolución mediante selección natural (Bennett y Fisher 1930; Wright 1931; Haldane 1990). Posteriormente T. Dobzhansky, en su obra *Genética y el origen de las especies* (Dobzhansky 1937), unió definitivamente ambas disciplinas en lo que, junto a posteriores aportaciones en citología, sistemática, botánica, morfología, ecología y paleontología, terminó por constituir la Síntesis evolutiva moderna, acuñada así por J. Huxley (1942).

Las bases de la genética evolutiva, aún vigentes hoy en día, establecen que la evolución está dirigida por cinco fuerzas principales: la mutación, como

única fuerza capaz de introducir nuevas variantes; la recombinación, generando nuevas combinaciones de variantes preexistentes; la deriva genética, aportando un componente estocástico al destino de dichas variantes; la migración o flujo genético, contrarrestando la diferenciación genética entre poblaciones y, finalmente, la selección natural, la única de las cinco fuerzas capaz de explicar la adaptación de las especies al medio. Todas estas fuerzas actúan asociadas a cambios demográficos, como expansiones o cuellos de botella, junto con los que determinan la dinámica evolutiva de las poblaciones.

No obstante, faltaba solventar un último problema para pasar del estudio de la evolución biológica al nacimiento de la evolución molecular: identificar la naturaleza del material hereditario y determinar la unidad mínima de cambio. Durante la primera mitad del siglo XX los estudios evolutivos se basaban en la variación de las características fenotípicas, ya fueran morfológicas, fisiológicas, etológicas o ecológicas, sin tener en cuenta la causa subyacente de dichas características. Fue a partir de la identificación del ADN como material hereditario (Avery, Macleod y McCarty 1944; Hershey y Chase 1952), la determinación de su estructura (Franklin y Gosling 1953; Watson y Crick 1953) y finalmente el desarrollo de las técnicas de secuenciación molecular tanto de proteínas (Sanger y Tuppy 1951) como de ADN (Sanger, Nicklen y Coulson 1977), cuando pudo empezar a estudiarse la evolución en base a los cambios aminoacídicos y nucleotídicos, es decir, analizando la variación genética causante de los cambios fenotípicos.

Teoría neutralista y evolución a diferentes escalas

Identificada la variación nucleotídica como la diana sobre la que actúan las distintas fuerzas evolutivas, faltaba resolver las normas que rigen el proceso. ¿Todos los cambios nucleotídicos tienen el mismo efecto?, ¿les afectan indistintamente las diferentes fuerzas evolutivas? La respuesta vino de la mano de la teoría neutralista. En 1968 Motoo Kimura propuso que la mayor parte de la variación observada a nivel molecular es selectivamente neutra, es decir, no disminuye ni aumenta la eficacia biológica y, por tanto, no es susceptible de verse afectada por la selección natural (Kimura 1968). De este modo, el destino de los cambios neutros depende únicamente de la deriva genética. Esta variación neutra puede localizarse, por ejemplo, en aquellas regiones intergénicas que no presenten elementos reguladores o, dentro de los propios genes, en los intrones o en las posiciones sinónimas de los codones, aquellas que no comportan un cambio de aminoácido dada la naturaleza degenerada del código genético (Figura 1).

Introducción

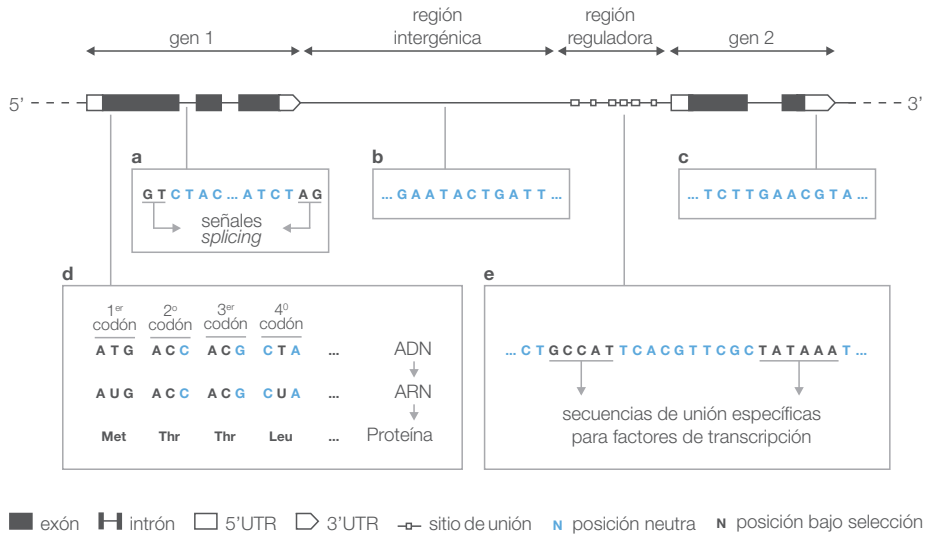


Figura 1. Ubicación de posibles posiciones nucleotídicas selectivamente neutras a lo largo de una región genómica. En la parte superior está representada una región compuesta por dos genes, separados por la región intergénica. a) En los intrones una gran mayoría de las posiciones son neutras. Una excepción la constituyen las posiciones que afectan a las señales de *splicing* GT y AG que los delimitan. b) y c) Tanto en las regiones intergénicas como en los UTR (*untranslated region*) 5' y 3' la inmensa mayoría de las posiciones son neutras. d) En las secuencias codificantes ubicadas en los exones las posiciones serán con mayor probabilidad neutras si son sinónimas. En el primer codón todas las posiciones presentan constricciones selectivas ya que cualquier cambio de nucleótido comportaría un cambio de aminoácido. En cambio, en el segundo y tercer codón la tercera posición es sinónima y por tanto posiblemente neutra, ya que ambos codones difieren en esta tercera posición y sin embargo codifican para el mismo aminoácido. e) En las regiones reguladoras la mayoría de las posiciones serán neutras exceptuando aquellas que forman parte de las secuencias funcionales capaces de reclutar a factores de transcripción que regularán la expresión de los genes adyacentes.

Otra fracción de los cambios nucleotídicos presenta un efecto deletéreo, disminuyendo la eficacia biológica de sus portadores. Estos cambios se ven afectados por la selección purificadora y normalmente son eliminados de la población, por lo que apenas contribuyen a la variación nucleotídica. Finalmente, la última fracción de variación corresponde a aquellos cambios que aumentan la eficacia biológica del individuo en un contexto ambiental determinado. Estos cambios están sujetos a la acción de la selección positiva, la cual hará que incrementen su frecuencia en la población. Que estos cambios lleguen a fijarse, alcanzando una frecuencia del 100%, dependerá del tipo de selección y de su intensidad, así como del tamaño de la población. Cuando la selección favorece a una única variante hablaremos de selección direccional y dicha variante podrá incrementar su frecuencia y llegar a la fijación, mientras que si la selección favorece a dos variantes en una misma posición nucleotídica hablaremos de selección balanceadora, alcanzándose frecuencias de equilibrio. Por otra parte, si la intensidad de la selección es baja,

en poblaciones pequeñas la deriva podría contrarrestar su efecto. También debe tenerse en cuenta cómo afecta la selección de una posición nucleotídica a las posiciones variables cercanas. La variación neutra ligada, por proximidad y/o ausencia de recombinación, a una variante seleccionada acabará compartiendo su mismo destino, lo que se denomina efecto de barrido selectivo. De manera similar, la selección positiva sobre una variante nucleotídica podría impedir que otra variante cercana pero seleccionada negativamente acabe siendo eliminada y viceversa (Figura 2).

Expuestos los mecanismos que determinan como actúan las fuerzas evolutivas sobre la variabilidad a nivel molecular, resta escoger la escala temporal a la que analizar el proceso. Por un lado, tenemos los estudios de polimorfismo enfocados en la variación intraespecífica y, por el otro, los estudios de divergencia centrados en la variación interespecífica. En el primer caso se analiza la evolución dentro de una misma especie y, por tanto, en una escala de tiempo breve y cercana. Desde la perspectiva de la teoría neutralista, el polimorfismo constituye una fase transitoria de la evolución en la que se puede observar el proceso mientras transcurre. Así pues, analizando la variabilidad intraespecífica se podrían detectar, por ejemplo, cambios beneficiosos a frecuencias intermedias que la selección aún no habrá tenido tiempo de fijar, huellas de la historia demográfica de la especie o, incluso, procesos de diferenciación entre poblaciones que podrían desembocar en eventos de especiación. Por el contrario, en los estudios de divergencia se analiza la evolución molecular entre distintas especies, géneros, familias u otros taxones superiores. Los estudios de divergencia acostumbran a utilizar un solo individuo como representante de cada especie, por lo que aportan principalmente información sobre los cambios fijados durante la divergencia de las especies analizadas. Sin embargo, esto nos permitirá estudiar la evolución a escalas de tiempo mucho más lejanas y extensas.

Tras la huella de la selección

Identificar aquellas variantes moleculares que suponen una mejor adaptación al medio y son diana de la selección positiva ha sido siempre uno de los objetivos principales del estudio de la evolución molecular, ya sea en referencia a posiciones nucleotídicas concretas o al conjunto de cambios que determinan, por ejemplo, una variante génica o alelo. Las aproximaciones para conseguirlo son distintas en función de si se analizan datos de polimorfismo o de divergencia, pero todas tienen en común la comparación de dos clases de posiciones: una focal y una

Introducción



Figura 2. Efecto de los distintos tipos de selección. Cada alineamiento consta de 6 secuencias representativas de la variabilidad nucleotídica de una población hipotética antes (Sec. 1-6) y después (Sec. 1'-6') de actuar la selección. a) La selección purificadora elimina de la población las variantes deletéreas presentes en las secuencias 2 y 6, así como la variante neutra ligada en la secuencia 2. b) La selección direccional termina fijando en la población la variante beneficiosa presente en la secuencia 5 y la variante neutra ligada, alcanzando ambas una frecuencia del 100% (barrido selectivo). c) La selección balanceadora desemboca en frecuencias de equilibrio para las variantes beneficiosas presentes en la misma posición nucleotídica en las secuencias 1 y 4.

neutra. En la clase focal se engloban aquellas posiciones candidatas o susceptibles de estar bajo el efecto de la selección positiva. Por contra, la clase neutra agrupa todas las posiciones donde, o bien las mutaciones no tengan consecuencias fenotípicas y por tanto sean invisibles para la selección natural, o bien los patrones de variación se ajusten al promedio de los genomas completos, caracterizados por estar moldeados principalmente por la deriva, la selección purificadora y los cambios demográficos.

Los métodos para detectar selección en estudios de polimorfismo se basan principalmente en los llamados test de neutralismo. La teoría neutralista determina que, en ausencia de selección, una población de tamaño constante alcanzará un equilibrio entre la aparición de nuevas variantes por mutación y la fijación/eliminación de éstas por deriva. Este modelo de equilibrio mutación-deriva es utilizado como hipótesis nula a la hora de contrastar los datos observados. Los distintos test de neutralismo, ya estén basados en los espectros de frecuencias (Tajima 1989; Fu y Li 1993; Fay y Wu 2000) o en la distribución de haplotipos (Fu 1997), están diseñados para identificar desviaciones del neutralismo.

No obstante, no solo la selección natural provoca estas desviaciones del neutralismo, también lo hacen los cambios demográficos de las poblaciones estudiadas. Por ejemplo, la selección positiva puede acabar fijando una variante beneficiosa y provocando un barrido selectivo, disminuyendo la variabilidad nucleotídica en la región. Posteriormente, nuevas mutaciones surgirán en la población, dando como resultado un patrón de variabilidad nucleotídica caracterizado por un exceso de variantes a baja frecuencia. Sin embargo, este patrón es exactamente el mismo que se encontraría tras un proceso demográfico de cuello de botella y posterior expansión poblacional.

Partiendo de esta base, ¿cómo se discrimina entre ambas causas? Si se trata de un proceso selectivo, el exceso de variantes a baja frecuencia estará acotado a la región próxima a la variante seleccionada. Si se trata de un proceso demográfico, dicho patrón se hallará extendido a todo el genoma. Aquí entra en juego la importancia de contrastar las clases focal y neutra mencionadas anteriormente. La clase focal puede ser una región candidata a estar bajo selección y la clase neutra una región de neutralidad ya contrastada. De este modo, si ambas regiones presentan la misma desviación del equilibrio mutación-deriva significa que ésta se debe a la historia demográfica de la población. Cambiando la perspectiva, también se podría considerar como clase neutra a un conjunto de genes, regiones genómicas o incluso la totalidad del genoma. En este caso, la clase focal no estaría definida *a priori* sino que se hallaría de forma prospectiva entre aquellas regiones que se desvíen del patrón de variabilidad promedio.

Sin abandonar los estudios de polimorfismo, existen métodos de detección de selección positiva que requieren información de la divergencia respecto a alguna especie próxima con el objetivo de contextualizar la variación intraespecífica. Estos métodos se basan en la premisa de que la relación polimorfismo-divergencia se mantiene constante bajo condiciones de neutralismo. De este modo, la proporción entre variantes polimórficas para una especie y variantes fijadas respecto a una segunda especie son comparadas entre las clases focal y neutra, siendo las posiciones que se adjudican a cada clase las que difieren según el método.

En el test HKA (Hudson, Kreitman y Aguadé 1987), al igual que en los test basados únicamente en variación intraespecífica, se comparan dos o más regiones, de modo que la clase neutra sea una región de neutralidad conocida o un conjunto de regiones que reflejen la relación polimorfismo-divergencia promedio en el genoma. Por el contrario, el test MK (McDonald y Kreitman 1991) puede aplicarse a una única región siempre que ésta sea codificadora, ya que

discrimina entre los cambios nucleotídicos sinónimos y no sinónimos dentro del propio gen. Los cambios sinónimos son utilizados como clase neutra, mientras que la clase focal está constituida por los cambios no sinónimos, aquellos que sí implican un cambio de aminoácido que puede modificar la actividad de la proteína resultante y, por tanto, están sujetos a la selección natural.

Si nos movemos al ámbito de los estudios de divergencia, el principal método para detectar selección positiva está basado, al igual que el test MK, en la comparación de la variación en posiciones sinónimas y no sinónimas, que actuarán como clase neutra y focal, respectivamente. En este caso, sin embargo, solo se dispone de una secuencia por especie, por lo que no se puede comparar variación intra e interespecífica y se asume que todos los cambios están fijados en las respectivas especies. Así, la búsqueda de selección se basará únicamente en la ratio entre la divergencia no sinónima y sinónima, o ω ($\omega = d_n/d_s$) (Miyata y Yasunaga 1980). En un gen que evolucione bajo el modelo neutro, principalmente pseudogenes (genes que han perdido su función), se observarán los mismos niveles de divergencia en posiciones sinónimas y no sinónimas, $\omega = 1$, ya que ningún cambio tendrá efectos fenotípicos. En cambio, en aquellos genes donde predomine la selección purificadora la mayor parte de los cambios no sinónimos habrán sido eliminados por ésta, por lo que $\omega < 1$. Finalmente, en los genes donde la selección positiva haya favorecido la fijación de cambios no sinónimos obtendremos que $\omega > 1$.

No obstante, a efectos prácticos apenas se identifican genes con una $\omega > 1$. Esto se debe a que si se analiza la divergencia de un grupo de especies es probable que solo una de ellas, o el ancestro de un subgrupo de ellas, haya sufrido el efecto de la selección positiva. Desde otro ángulo, puede que la selección haya actuado sobre todas las especies pero únicamente en una pequeña región del gen. En ambos casos la selección purificadora generalizada enmascarará la selección positiva puntual, arrojando por tanto una ω que continuará siendo < 1 . Para solucionar este problema se han desarrollado métodos que, partiendo de la reconstrucción filogenética de las especies estudiadas y utilizando un marco estadístico de máxima verosimilitud, son capaces de estimar una ω para cada rama del árbol filogenético y para cada codón del gen, permitiendo así identificar episodios puntuales de selección positiva (Nielsen y Yang 1998; Yang y Nielsen 1998; Yang y Swanson 2002) (Figura 3).

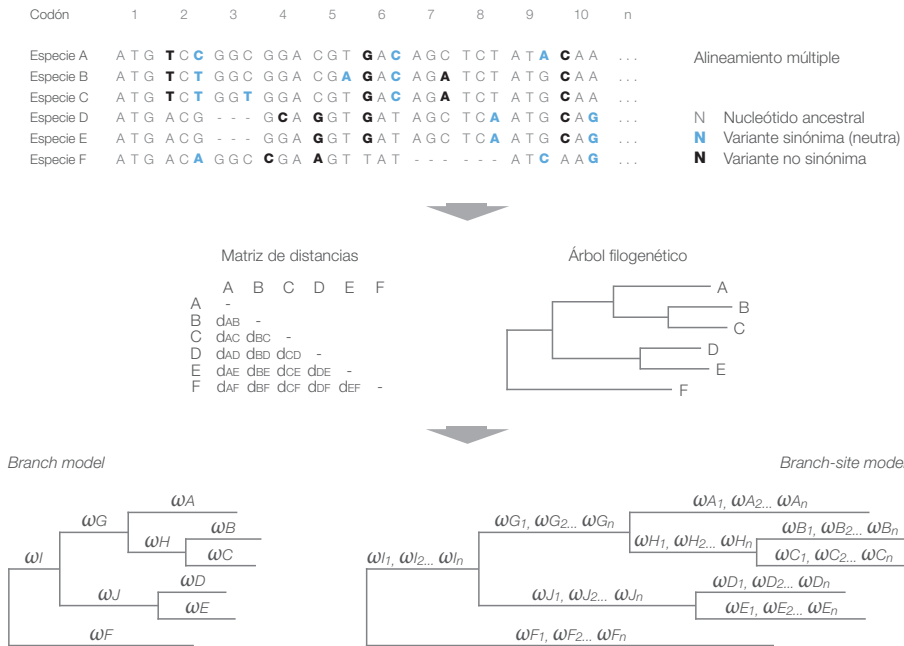


Figura 3. Proceso de reconstrucción filogenética y detección de selección en estudios de divergencia de regiones codificantes. Partiendo de las secuencias nucleotídicas de diversas especies se realiza un alineamiento múltiple, incluyendo *gaps* (-) donde haya inserciones o deleciones con tal de mantener la homología en todas las posiciones. A partir del alineamiento múltiple se puede reconstruir el árbol filogenético de las especies. Si se aplica un método de reconstrucción basado en distancias genéticas, en primer lugar se tendrán que estimar dichas distancias (*d*) entre cada par de especies, transformando, mediante modelos de acumulación de cambios, el número de diferencias observadas en el alineamiento en el número de sustituciones reales. El árbol filogenético obtenido permite visualizar las relaciones evolutivas entre las especies y la longitud de cada rama, incluyendo las internas para los correspondientes ancestros. Finalmente, los métodos de detección de selección por máxima verosimilitud permiten calcular las ω (d_N/d_S) para cada rama del árbol (*Branch model*) o incluso para cada codón de cada rama independientemente (*Branch-site model*). Ambos modelos admiten variaciones que permitirán realizar contrastes de hipótesis y detectar episodios puntuales de selección ($\omega > 1$).

Distintas aproximaciones: del gen al genoma

Los primeros estudios de evolución molecular basados en secuencias nucleotídicas se enfocaron en analizar un solo gen a la vez. Uno de los primeros y más importantes ejemplos sería el conjunto de trabajos sobre el gen *Adh* en la mosca del vinagre *Drosophila melanogaster* (Kreitman 1983; Hudson, Kreitman y Aguadé 1987; Kreitman y Hudson 1991) y posteriormente en la planta herbácea *Arabidopsis thaliana* (Hanfstingl et al. 1994). No obstante, los métodos de secuenciación han ido perfeccionándose y abaratándose ininterrumpidamente hasta el día de hoy. Actualmente, gracias a las técnicas de secuenciación masiva, es posible estudiar

la evolución molecular de genomas completos, ya sea en organismos modelo como *Drosophila* (Clark et al. 2007) o la especie humana (Auton et al. 2015), o en organismos no modelo pero de valor científico, social o incluso económico, como el café (Denoeud et al. 2014).

Durante este tránsito entre los estudios pioneros sobre un solo gen hasta el análisis de genomas completos, el campo de la evolución molecular en las últimas décadas se ha centrado principalmente en un punto intermedio, es decir, trabajos que abarcan un conjunto de genes o regiones genómicas. Estos estudios, denominados multilocus, pueden dividirse en dos clases en función del objetivo del estudio. En primer lugar, como se ha mencionado anteriormente, pueden utilizarse un conjunto de regiones, generalmente no codificadoras, como representación del patrón de variabilidad neutra característico de una determinada especie o población, lo cual permite inferir su historia demográfica (Pratdesaba, Segarra y Aguadé 2015). A partir de este patrón se pueden identificar genes *outliers* que presenten desviaciones por el efecto de la selección adaptativa.

En segundo lugar, existe la opción de estudiar un grupo de genes relacionados funcionalmente. De hecho, los genes, así como las proteínas codificadas por estos, rara vez actúan individualmente sino que controlan los procesos biológicos de forma coordinada, por lo que resulta útil estudiarlos en conjunto. De este modo se han realizado estudios sobre genes implicados en una misma ruta metabólica, como la vía de la insulina en *Drosophila* (Guirao-Rico y Aguadé 2011; Alvarez-Ponce et al. 2012) o la vía de los fenilpropanoides en *A. thaliana* (Ramos-Onsins et al. 2008); genes que de forma independiente participan en los mismos procesos biológicos, como la resistencia a la salinidad en *A. thaliana* (Puerma y Aguade 2013); o genes cuyas proteínas, además de participar en un mismo proceso, lo hacen interactuando físicamente formando complejos multiprotéicos, como los genes reguladores Polycomb, cuya evolución molecular dentro del género *Drosophila* será el objeto de estudio de la presente tesis doctoral.

Los genes del grupo Polycomb

Función de las proteínas Polycomb

Las proteínas codificadas por los genes del grupo Polycomb constituyen un sistema de silenciamiento transcripcional presente en todos los grandes grupos de eucariotas, desde organismos unicelulares hasta humanos, siendo el díptero *D. melanogaster* la especie en la que han sido estudiados con mayor detalle (Simon y Kingston 2009; Lanzuolo y Orlando 2012). Las proteínas Polycomb actúan como reguladores epigenéticos, modificando residuos específicos de lisina en las histonas. En función de estas modificaciones, la cromatina adoptará una conformación más laxa o más compacta, variando de este modo la accesibilidad al ADN de la maquinaria transcripcional (Figura 4). Así, las proteínas Polycomb modifican las histonas presentes en las regiones de los genes a silenciar, induciendo la compactación de la cromatina e impidiendo que dichos genes se expresen. Además, con la participación de otros sistemas reguladores (Van Bortle et al. 2012), las proteínas Polycomb son capaces de modificar la estructura tridimensional de la cromatina, uniendo y silenciando simultáneamente regiones distantes del genoma (Bantignies y Cavalli 2011).

La inhibición de la transcripción por parte de las proteínas del grupo Polycomb tiene su contrapartida en la capacidad activadora de las proteínas del grupo Trithorax. Éstas últimas también realizan modificaciones en las histonas, pero en este caso su función es impedir la acción de las proteínas Polycomb y, por tanto, mantener la cromatina en una conformación extendida que facilite la expresión de los genes afectados (Schuettengruber et al. 2011; Steffen y Ringrose 2014). Cabe señalar que este sistema de regulación mediado por las proteínas Polycomb y Trithorax no es el que decide qué genes serán expresados, sino que cumple una función de mantenimiento de estados transcripcionales predefinidos. Un buen ejemplo de ello es el de los genes Hox, claves en el desarrollo embrionario y una de las dianas de las proteínas del grupo Polycomb mejor estudiadas. Cada

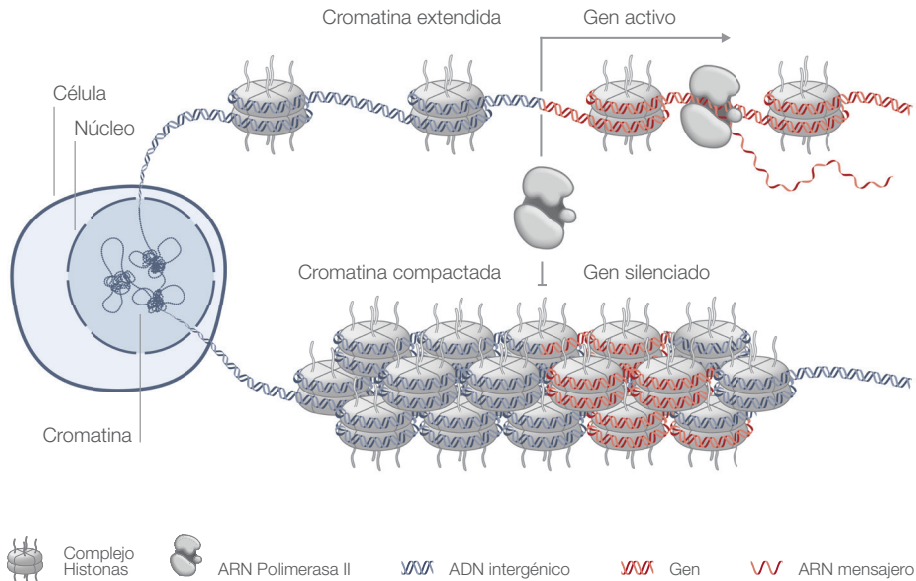


Figura 4. Estados de la cromatina. Dentro del núcleo, el ADN se organiza en fibras de cromatina, formadas por el ADN, las histonas y otras proteínas. Si la cromatina adopta una conformación extendida, como por ejemplo, tras la acción de las proteínas del grupo Trithorax, la maquinaria transcripcional puede acceder a los genes e iniciar la transcripción de los mismos, sintetizando el ARN mensajero que a su vez será traducido a las correspondientes proteínas. Si por el contrario la cromatina se encuentra compactada, como sucede tras la regulación mediada por las proteínas del grupo Polycomb, ni la ARN polimerasa II ni el resto de proteínas que integran la maquinaria transcripcional pueden acceder a las secuencias de los genes ni a sus elementos promotores y reguladores, por lo que dichos genes se mantienen inactivos.

uno de los distintos genes Hox es el encargado de coordinar la formación de las estructuras adecuadas en los diferentes segmentos del embrión (Hughes y Kaufman 2002; Pick y Heffer 2012). Por tanto, la expresión de estos genes debe estar sujeta a una regulación precisa que permita que cada uno de ellos se exprese en el segmento y momento adecuado. De esta regulación se encargan los factores transcripcionales de segmentación (Peel, Chipman y Akam 2005). No obstante, dichos factores únicamente están activos durante los estadios iniciales del desarrollo, tras los cuales entran en juego las proteínas de los grupos Polycomb y Trithorax, manteniendo, para cada gen Hox y en cada segmento embrionario, el estado transcripcional silenciado o activo previamente definido por los factores de segmentación.

Los genes Hox no son los únicos silenciados por las proteínas Polycomb. Estudios abarcando la totalidad del genoma de *D. melanogaster* identificaron más de 200 regiones sujetas a su actividad reguladora, principalmente genes que codifican para factores de transcripción y componentes de vías de señalización

(Schuettengruber et al. 2009). Este elevado número de dianas es consistente con la gran variedad de procesos biológicos en los que se ha detectado la implicación de las proteínas del grupo Polycomb tanto en animales como en plantas: inactivación del cromosoma X (Payer y Lee 2008), impronta genética (Wu y Bernstein 2008), prevención de la senescencia (Dietrich et al. 2007), regulación de células madre (Pietersen y Van Lohuizen 2008), control de la floración (Chanvivattana et al. 2004) o cáncer (Koppens y Van Lohuizen 2015), entre otros.

Los complejos Polycomb

El grupo Polycomb en el género *Drosophila* se compone de al menos 19 genes cuyas proteínas actúan formando diversos complejos multiméricos, cada uno de ellos con un papel diferenciado dentro del proceso de silenciamiento génico. Se han identificado hasta la fecha 5 complejos proteicos: PRC2 (*Polycomb repressive complex 2*), PRC1 (*Polycomb repressive complex 1*), dRAF (*dRING-associated factors*), PR-DUB (*Polycomb repressive deubiquitinase*) y PhoRC (*Pho repressive complex*) (Figura 5).

El complejo PRC2, compuesto por 4 proteínas, actúa mono- y dimetilando la lisina 27 de la histona H3 (H3K27me/me₂), lo que supone una señal para la compactación de la cromatina y por tanto su silenciamiento (Nekrasov, Wild y Müller 2005; Müller y Verrijzer 2009). PRC2 actúa ubicuamente y no se focaliza en una serie de genes diana, por lo que se cree que su función es la de prevenir la expresión inespecífica a lo largo del genoma (Lee et al. 2015). En este sentido, cabe señalar que la metilación de esta lisina (H3K27me) es mutuamente excluyente respecto a su acetilación activadora (H3K27ac) por parte de los complejos Trithorax. Enhancer of zeste (E(Z)) es la unidad catalítica encargada de la metilación (Czermin et al. 2002), mientras que Extra sex combs (ESC) (Tie et al. 2007) o, alternativamente, Esc-like (ESCL) (Kurzahls et al. 2008) proporciona la especificidad y facilita la unión a la histona H3. Por su parte, Suppressor of zeste 12 (SU(Z)12) (Chen, Birve y Rasmuson-Lestander 2008) y Chromatin assembly factor 1 subunit p55 (CAF1-55, también llamada NURF-55) (Wen, Quan y Xi 2012) resultan importantes para la unión al conjunto de histonas o nucleosoma. El complejo PRC2 puede ampliarse con una quinta proteína, Polycomb-like (PCL), formando el complejo alternativo Pcl-PRC2 que presenta un incremento de la actividad de E(Z), siendo así capaz de trimetilar la lisina 27 de la histona H3 (H3K27me₃) específicamente en las secuencias reguladoras de los genes diana, lo cual concentra y aumenta la señal silenciadora (Nekrasov et al. 2007).

Introducción

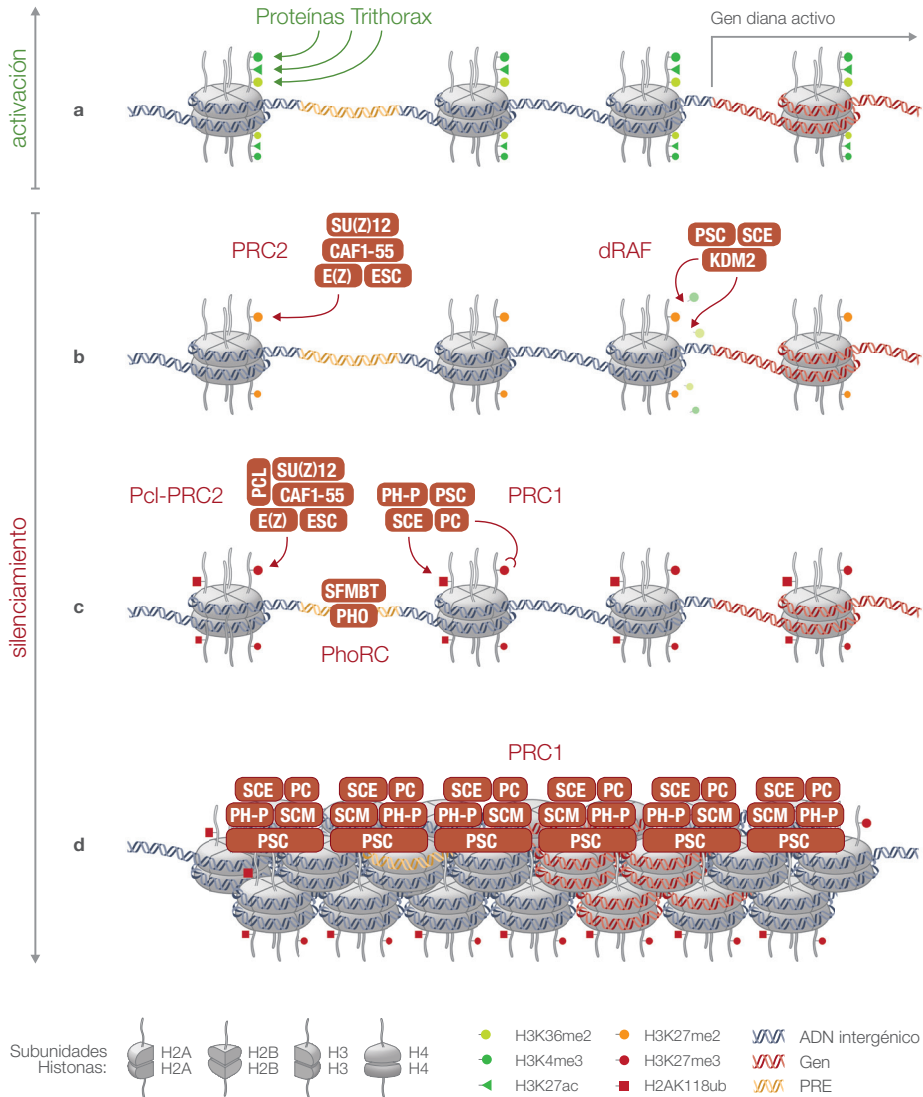


Figura 5. Mecanismos de acción de los complejos Polycomb. a) Las proteínas Trithorax introducen las señales activadoras de la transcripción. Una vez que estas proteínas han acetilado la lisina 27 de la histona H3 (H3K27ac) la activación del gen diana es irreversible. b) Por el contrario, las marcas activadoras H3K4me3 y H3K36me2 sí pueden ser eliminadas por el complejo dRAF, lo que a su vez permite que el complejo PRC2 dimetile la lisina 27 de la histona H3 (H3K27me2), impidiendo su acetilación e iniciando el proceso de silenciamiento génico. c) El complejo PhoRC se une a las secuencias PRE (*Polycomb response elements*), presentes en la región reguladora de los genes diana, reclutando a los complejos Pcl-PRC2 y PRC1. El primero trimetila la lisina 27 de la histona H3 previamente dimetilada (H3K27me3). El segundo, tras reconocer la marca H3K27me3, monoubiquitina la lisina 118 de la histona H2A (H2AK118ub). d) Finalmente, con la ayuda de otros complejos remodeladores de la cromatina, los complejos PRC1 inician la compactación de la región génica mediante la capacidad polimerizadora de PSC, PH-P y, probablemente, SCM.

El complejo PRC1 se compone de otras 4 proteínas y es capaz de reconocer las marcas dejadas por el complejo Pcl-PRC2 para, a continuación, modificar las histonas con su propia señal silenciadora que consiste en la monoubiquitinación de la lisina 118 de la histona H2A (H2AK118ub) (Shao et al. 1999; Müller y Verrijzer 2009). La proteína Polycomb (PC) fue la primera en identificarse, dando nombre al grupo, y es la encargada de reconocer la señal H3K27me3 (Fischle et al. 2003), localizando el objetivo para la proteína Sex combs extra (SCE, también llamada dRING), la cual presenta la actividad ubiquitina ligasa (Gutierrez et al. 2012). La proteína Posterior sex combs (PSC) o, alternativamente, su homóloga Suppressor of zeste 2 (SU(Z)2) (Lo, Ahuja y Francis 2009), también resulta imprescindible para el proceso de silenciamiento génico gracias a su capacidad para hacer de puente entre los nucleosomas, inhibiendo así la remodelación de la cromatina (Lo, McElroy y Francis 2012). Por su parte, el papel de Polyhomeotic proximal (PH-P) aún no se conoce en profundidad aunque se especula que puede ser similar al de PSC dada su capacidad para formar polímeros (Robinson et al. 2012). PH-P también presenta una proteína homóloga, Polyhomeotic distal (PH-D) (Beisswanger y Stephan 2008), aunque ésta solo se encuentra en un subconjunto de las especies de *Drosophila*. Finalmente, el complejo PRC1 incluye en ocasiones una quinta proteína, Sex comb on midleg (SCM), que contribuiría a la actividad polimerizadora de PH-P (Peterson et al. 1997).

SCE y PSC también forman parte del complejo funcional dRAF, que incluye además la proteína Lysine (K)-specific demethylase 2 (KDM2) (Lagarou et al. 2008; Müller y Verrijzer 2009). Esta tercera subunidad potencia la función ubiquitina ligasa de SCE y, mediante su actividad desmetilasa, es capaz de eliminar las marcas activadoras H3K4me3 y H3K36me2 (Kavi y Birchler 2009). De este modo, el complejo dRAF no solo contribuye a incorporar señales inhibitoras de la transcripción, sino que además elimina las señales activadoras dejadas por las proteínas Trithorax. En cuanto al complejo PR-DUB (Scheuermann et al. 2010), último en ser descubierto, está formado únicamente por dos proteínas, Additional sex combs (ASX) (Sinclair et al. 1998) y Calypso (Gaytán de Ayala Alonso et al. 2007). Aunque se sabe que ASX resulta imprescindible para el correcto funcionamiento del complejo PR-DUB, aún no se ha descifrado su función. Calypso presenta actividad deubiquitinasa específica para la histona H2A, aunque por el momento no se conoce la posición sobre la que actúa (Scheuermann, Gutiérrez y Müller 2012).

Finalmente, el complejo PhoRC es el encargado de identificar los genes diana y reclutar al resto de complejos (Klymenko et al. 2006; Müller y Verrijzer 2009). Está compuesto por dos proteínas, Pleiohomeotic (PHO) o,

alternativamente, su homóloga Pho-like (PHOL) y la proteína codificada por el gen *Scm-related gene containing four MBT domains* (SFMBT). PHO y PHOL son las dos únicas proteínas Polycomb capaces de unirse al ADN, concretamente a unas secuencias específicas denominadas *Polycomb response elements* (PRE) situadas en la región reguladora de los genes diana (Fritsch et al. 1999; Brown et al. 2003). Por su parte, SFMBT cumple una función estabilizadora gracias a su capacidad para reconocer y unirse a diversos residuos metilados en las histonas H3 y H4 (Grimm et al. 2009).

El modelo clásico para explicar la movilización de los distintos complejos Polycomb en los genes diana sostiene que, tras el reconocimiento de las secuencias PRE por parte de PhoRC, éste recluta a PRC2 y Pcl-PRC2 mediante la interacción directa entre PHO y la pareja E(Z)-ESC (Wang et al. 2004), para que posteriormente PRC1 reconozca la señal H3K27me3 dejada por Pcl-PRC2 en las histonas adyacentes. No obstante, se ha propuesto que PhoRC podría reclutar directamente a PRC1 mediante interacciones entre PHO y PC (Mohd-Sarip et al. 2002), así como entre SFMBT y SCM (Wang et al. 2010). Por otra parte, se ha sugerido que PRC1 podría ser necesario para la correcta unión de PhoRC a las secuencias PRE (Kahn et al. 2014), o incluso que el reclutamiento de PRC1 y PRC2 podría ser independiente de las señales dejadas por ambos en las histonas (Kahn et al. 2016). Respecto a dRAF y PR-DUB, se desconoce cómo consiguen focalizar su acción en los genes a silenciar.

Polycomb response elements

Las secuencias PRE, presentes en las regiones reguladoras de los genes silenciados por las proteínas Polycomb, están constituidas por un conjunto de sitios de unión para diversos factores de transcripción (Brown y Kassis 2013; Kassis y Brown 2013). Naturalmente, el sitio de unión para las proteínas PHO y PHOL es uno de ellos y consiste en variaciones de la secuencia nucleotídica GCCAT (Cunningham, Brown y Kassis 2010). No obstante, ni PHO ni PHOL son suficientes por sí mismas para reclutar los distintos complejos Polycomb. Para ello es necesaria la participación de un conjunto de cofactores, entre los que destacan las proteínas Trithorax-like (TRL) (Mahmoudi et al. 2003) y Zeste (Z) (Hur et al. 2002). Cabe destacar la naturaleza bivalente de las secuencias PRE, ya que no solo se ha observado su asociación con los complejos Polycomb sino que también podrían intervenir en el reclutamiento de las proteínas Trithorax, actuando por tanto como activadores o silenciadores en distintas circunstancias.

No se conoce con exactitud cómo se produce el cambio entre PRE y TRE (*Trithorax response elements*), aunque se cree que su propia transcripción en forma de ARN no codificante podría estar implicada (Bender y Fitzgerald 2002; Herzog et al. 2014)

No existe una secuencia consenso que permita identificar fácilmente estos elementos reguladores, ya que cada PRE estudiado hasta la fecha es distinto en cuanto al número y orden de los distintos sitios de unión que lo componen. De hecho, se ha observado que su composición varía en función de las necesidades transcripcionales del gen regulado (Okulski et al. 2011), por lo que no serían intercambiables. Esta composición variable condiciona a su vez la evolución de los PRE, de modo que las constricciones evolutivas que los afectan no están dirigidas a la conservación de su secuencia completa sino al mantenimiento del número necesario de sitios de unión y su orden adecuado, siendo indiferente que dichos sitios de unión sean homólogos entre especies o surjan *de novo* (Hauenschild et al. 2008).

Drosophila como organismo de estudio

El género *Drosophila*

El género *Drosophila* pertenece a la familia Drosophilidae del orden Díptera y se compone de más de 2000 especies. Son insectos holometábolos de pequeño tamaño, entre 2 y 8 mm, que viven y se reproducen principalmente sobre material vegetal en descomposición, nutriéndose de los hongos, levaduras y bacterias que lo fermentan. Especies como *D. melanogaster*, comúnmente llamada mosca de la fruta o del vinagre, llevan más de un siglo siendo utilizadas como organismo modelo en campos de estudio como la ecología, la biología molecular o la genética (Hales et al. 2015). Son varias las razones que convierten a estos insectos en una herramienta científica extremadamente útil. Su dieta y pequeño tamaño facilitan su mantenimiento en el laboratorio, mientras que su breve tiempo de generación, entre 10 días y un mes, permite estudiar sucesivas generaciones en un periodo de tiempo relativamente corto. Dan lugar a una descendencia abundante, algo imprescindible a la hora de contrastar hipótesis estadísticamente. Finalmente, como ocurre en otros dípteros, presentan cromosomas politénicos en sus glándulas salivales. Estos cromosomas están constituidos en realidad por la superposición de múltiples cromátidas interfásicas replicadas dentro de una misma célula. Esto permite que sean observables al microscopio óptico con gran resolución y facilita las técnicas de hibridación *in situ*.

El género *Drosophila* se divide principalmente en dos subgéneros, *Sophophora* y *Drosophila*. No obstante, el origen monofilético del subgénero *Drosophila* se ha puesto en duda, sugiriéndose que uno de los grupos que lo componen, las *Drosophilas* hawaianas, podrían constituir por si mismas un género denominado *Idiomyia* (Remsen y O'Grady 2002). Por su parte, el subgénero *Sophophora*, el más amplio y estudiado, se compone de cuatro grupos: *melanogaster*, *obscura*, *willistoni* y *saltans*.

Tras la pionera secuenciación de los genomas completos de *D. melanogaster* (Adams et al. 2000) y *D. pseudoobscura* (Richards et al. 2005) se sumaron, en el año 2007, los genomas de otras 10 especies de *Drosophila* por parte del *Drosophila 12 Genomes Consortium* (Clark et al. 2007). Estos 12 genomas, accesibles desde la base de datos FlyBase (<http://flybase.org>), representan a la mayor parte de los grupos dentro de la filogenia de *Drosophila* y constituyen probablemente la herramienta más importante en el campo de la genética evolutiva centrada en este organismo modelo. En los últimos años, los métodos de secuenciación masiva y otras técnicas de alto rendimiento han seguido avanzando, por lo que hoy en día el número de genomas de distintas especies de *Drosophila* secuenciados asciende a más de una veintena, las líneas secuenciadas dentro de la especie *D. melanogaster* se cuentan por cientos (Mackay et al. 2012) y han surgido bases de datos centradas en aspectos funcionales de los genomas más allá de sus secuencias, entre las que destaca el proyecto modENCODE (Roy et al. 2010).

El grupo *obscura* y *Drosophila subobscura*

El grupo *obscura*, perteneciente al subgénero *Sophophora* de *Drosophila*, está compuesto por unas 35 especies divididas en 4 subgrupos: *microlabis*, de distribución africana; *obscura*, presente principalmente en Europa y también llamado subgrupo *obscura* del Viejo Mundo; y los subgrupos *affinis* y *pseudoobscura*, localizados en el continente americano y denominados en conjunto *obscuras* del Nuevo Mundo. Dos especies del subgrupo *pseudoobscura*, *D. pseudoobscura* y *D. persimilis*, están entre las 12 secuenciadas por el *Drosophila 12 Genomes Consortium* y, junto con las otras 10, han sido utilizadas en la presente tesis doctoral. Además, la representación del grupo *obscura* se ha ampliado a otras 4 especies: por un lado *D. miranda*, incluida también en el subgrupo *pseudoobscura*, y por otro *D. guanche*, *D. madeirensis* y *D. subobscura*, pertenecientes al subgrupo *obscura* (Figura 6).

D. subobscura es de especial relevancia en este trabajo dado que es la especie elegida para analizar la variabilidad nucleotídica de los genes Polycomb a nivel intraespecífico. Es una especie de pigmentación oscura, generalista, polífaga y con un amplio rango de distribución por toda la región paleártica, desde el Sahara hasta el norte de Europa, a excepción de Islandia y parte de Escandinavia. También se encuentra en las islas Canarias, donde cohabita con la especie endémica *D. guanche* (Monclús 1976); en la isla de Madeira, donde cohabita con la especie también endémica *D. madeirensis* (Monclús 1984) y en las islas Azores. Recientemente ha colonizado el continente americano, extendiéndose a partir de los años 70 del

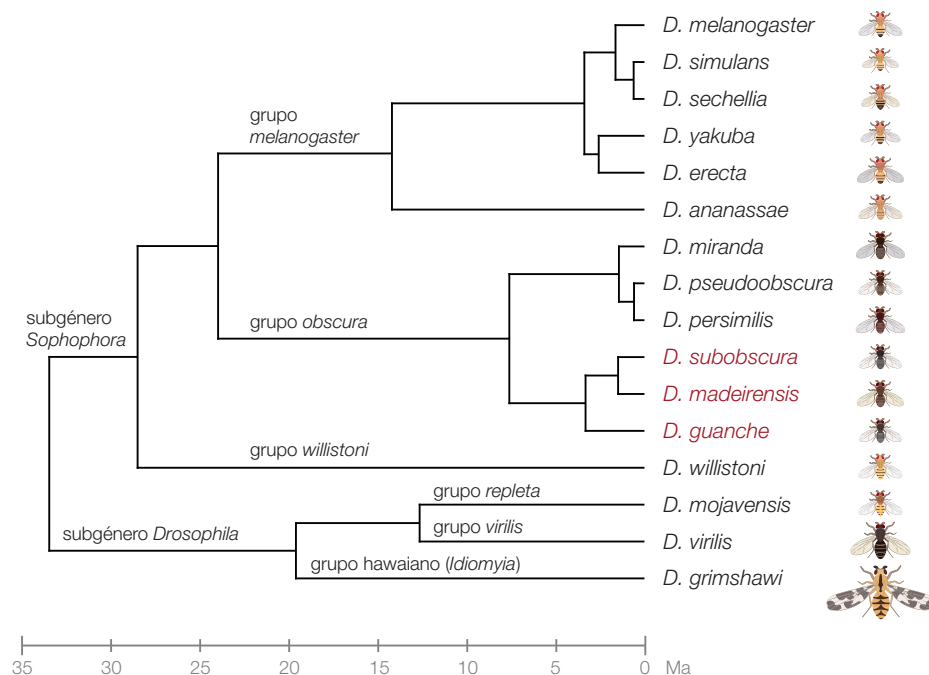


Figura 6. Filogenia de las especies de *Drosophila* estudiadas en la presente tesis doctoral. Las especies representadas en negro son aquellas cuyos genomas completos están disponibles en la base de datos FlyBase, las doce secuenciadas inicialmente por el *Drosophila* 12 Genomes Consortium (Clark et al. 2007) más *D. miranda*, añadida posteriormente (Zhou y Bachtrog 2012). Las secuencias de los genes Polycomb de las tres especies del grupo *obscura* representadas en rojo fueron obtenidas como parte del trabajo experimental de la presente tesis. Los tiempos de divergencia indicados en millones de años (Ma) provienen de las estimaciones de Obbard et al. (2012).

siglo pasado por la costa oeste tanto de Suramérica (Brneie et al. 1981) como de Norteamérica (Beckenbach y Prevosti 1986). En la naturaleza suele hallarse en los márgenes de los bosques de robles, pinos, abetos, castaños u olmos, desde el nivel del mar hasta una altura de más de 2000 metros. En condiciones óptimas de laboratorio y a 18° C su ciclo biológico es de aproximadamente tres semanas. Su densidad poblacional en la naturaleza varía con las estaciones de acuerdo a la temperatura, llegando prácticamente a desaparecer en invierno (Krimbas 1993). Su cariotipo consta de un par de cromosomas puntiformes y 5 pares de cromosomas acrocéntricos, denominados U, E, J, O y A, siendo este último el cromosoma sexual. La especie se caracteriza por presentar un elevado polimorfismo de inversiones cromosómicas en poblaciones naturales, habiéndose identificado más de 60 inversiones diferentes, agrupadas en más de 85 ordenaciones (Krimbas y Powell 1992).

Evolución de los genes Polycomb en *Drosophila*: consideraciones previas

Antes de iniciar el estudio evolutivo de un determinado grupo de genes cabe preguntarse qué dinámicas evolutivas se esperan de ellos en base a su origen, función u organización en los organismos analizados. Por ejemplo, en familias génicas implicadas en la interacción directa con el medio, como las que forman parte de los sistemas inmunológico o quimiosensorial, se esperarán tasas de cambio elevadas ya sea a nivel de variación aminoacídica (Foret y Maleszka 2006; Lazzaro 2008) o en cuanto a ganancia y pérdida de genes (Quesada, Ramos-Onsins y Aguadé 2005; Sánchez-Gracia, Vieira y Rozas 2009). Por el contrario, cabe suponer que genes como los Hox o los Polycomb, con una posición central en los sistemas reguladores y, por tanto, altamente pleiotrópicos, estén más conservados evolutivamente debido a la selección purificadora. No obstante, esto no implica ni mucho menos una estasis evolutiva o la ausencia de selección positiva. De hecho, aunque sus cambios adaptativos puedan ser menos frecuentes, sus consecuencias pueden ser más drásticas e implicar grandes saltos evolutivos (Pick y Heffer 2012).

Los genes Hox y su regulación por parte de los complejos Polycomb están implicados en diferencias morfológicas dentro del género *Drosophila*. Concretamente, el gen Hox *Sex combs reduced* (*Scr*) juega un papel clave en la variación del número y tamaño de los peines tarsales, estructuras presentes en el primer par de patas de los machos de algunas especies de *Drosophila* y cuya función es estimular a las hembras durante el apareamiento (Randsholt y Santamaria 2008; Tanaka et al. 2011). Las mutaciones que afectan a los genes Polycomb pueden producir transformaciones homeóticas consistentes en la sustitución de los segmentos meso y metatorácicos por segmentos protorácicos, lo que se traduce en la transformación del segundo y tercer par de patas en una repetición del primer par de patas y, por tanto, en la presencia de peines tarsales en todas las patas de los machos. Estas transformaciones son muy similares a las que se observan en los híbridos de las especies hermanas *D. subobscura* y *D. madeirensis*, de lo que se deduce que los genes Polycomb podrían estar implicados en el proceso de especiación entre ambas (Papaceit, San Antonio y Prevosti 1991).

El hecho de que las proteínas Polycomb actúen formando complejos multiméricos es otro aspecto que añade interés al estudio de su evolución molecular. Se ha observado que las proteínas que interactúan físicamente tienden a presentar tasas evolutivas correlacionadas (Hakes et al. 2007; Lovell y Robertson 2010), probablemente debido a que cambios en la secuencia aminoacídica de una proteína deberán ser compensados a su vez por cambios en las proteínas con las que hacen contacto, de modo que la interacción entre ellas siga siendo óptima. Por tanto, las proteínas del grupo Polycomb son claras candidatas a presentar cierto grado de coevolución, especialmente aquellas que integran un mismo complejo. Además, si estos cambios compensatorios suponen una ventaja para la función de estas proteínas, es de esperar que estén sujetos a la selección positiva.

Los genes del grupo Polycomb también se caracterizan por presentar entre sus integrantes diversos genes parálogos, esto es, originados tras un evento de duplicación. Los pares de genes *Pho* y *Phol*, *Esc* y *Escl* y *Ph-p* y *Ph-d* serían ejemplos de este proceso. Cuando un gen se duplica, la función de una de las dos copias pasa a ser redundante, por lo que es probable que la selección purificadora deje de actuar sobre ella, provocando que acumule cambios deletéreos hasta perderse en forma de pseudogen. No obstante, en algunos casos como los genes Polycomb mencionados, ambas copias pueden mantenerse tras sufrir un proceso de subfuncionalización, en el que ambas copias retienen un subconjunto de las funciones originales, o de neofuncionalización, en el caso de que una copia retenga la función ancestral mientras que la otra adquiere una función nueva (Assis y Bachtrog 2013). La identificación de una nueva duplicación de un gen Polycomb, durante los estadios preliminares de la presente tesis doctoral, nos brindó la oportunidad de analizar su evolución molecular y compararla con la del resto de genes Polycomb duplicados.

Finalmente, cabe contemplar la posibilidad de que la evolución de un grupo de genes reguladores como los Polycomb conlleve a su vez la evolución de los genes diana sobre los que ejercen su función, especialmente de sus regiones reguladoras. Además, la conservación a lo largo de las regiones reguladoras entre diversas especies es una herramienta especialmente útil a la hora de identificar elementos reguladores concretos. En base a estas premisas, se ha escogido el gen *Hox Scr* para realizar un análisis de su región reguladora a través de la filogenia de *Drosophila*. Este estudio comparativo puede contribuir a la identificación de secuencias reguladoras como las PRE y permite observar la evolución de estos elementos.

Objetivos

Los genes del grupo Polycomb tienen un gran interés genético y evolutivo, tanto por su papel clave en la regulación de la expresión génica, como por su posible participación en los procesos de especiación al controlar los patrones morfogénicos durante el desarrollo embrionario. Partiendo de esta premisa, el principal objetivo de la presente tesis doctoral es ofrecer por primera vez una visión amplia y detallada de la evolución molecular de los genes Polycomb en las especies del género *Drosophila*, empleando distintas aproximaciones pero con un denominador común: la detección de selección adaptativa. Para ello, se analizarán los patrones de variación molecular tanto a nivel interespecífico, en la filogenia de *Drosophila*, como a nivel intraespecífico, en la especie *D. subobscura*. A su vez, se estudiarán las consecuencias adaptativas de la actuación de las proteínas del grupo Polycomb formando complejos, de la duplicación de uno de sus miembros y de su acción sobre la región reguladora de uno de sus genes diana.

Los objetivos específicos de esta tesis doctoral son:

- Comparar los niveles y patrones de divergencia interespecífica de los distintos genes Polycomb en el género *Drosophila*, caracterizando el grado de conservación tanto a nivel global como a lo largo de sus dominios funcionales.
- Determinar si la divergencia de los genes Polycomb en el género *Drosophila* está afectada por un proceso de coevolución debido a las interacciones físicas entre las proteínas que forman cada uno de los distintos complejos funcionales.

- Detectar incrementos en la proporción de sustituciones no sinónimas respecto a las sinónimas en ramas concretas de la filogenia de *Drosophila* e inferir episodios puntuales de selección adaptativa, identificando las posiciones seleccionadas.
- Caracterizar el proceso evolutivo que tuvo lugar tras la duplicación del gen *Caf1-55* en las especies de *Drosophila* del grupo *obscura* para poder inferir cómo ha actuado la selección sobre la nueva copia (*Caf1-55dup*).
- Analizar los niveles y patrones de variación nucleotídica intraespecífica de los genes Polycomb en *D. subobscura* con la finalidad de detectar desviaciones del modelo neutro e inferir selección adaptativa reciente.
- Identificar y determinar la conservación evolutiva de los distintos elementos reguladores en *cis* presentes en la región *upstream* del gen *Scr*, diana de la regulación mediada por el grupo Polycomb y los factores de segmentación.

Informe de las
directoras

Las Dras. Carmen Segarra Robert y Montserrat Papaceit Vidal, directoras de la tesis doctoral realizada por el Sr. Juan Manuel Calvo Martín, con el título “Evolución molecular de los genes del grupo Polycomb en el género *Drosophila*”

INFORMAN

Que la presente tesis doctoral está elaborada como un compendio de cuatro artículos científicos con datos originales, tres de los cuales ya han sido publicados en revistas internacionales de reconocido prestigio:

1

Calvo-Martín JM, Librado P, Aguadé M, Papaceit M, Segarra C (2016) Adaptive selection and coevolution at the proteins of the Polycomb repressive complexes in *Drosophila*. *Heredity* 116:213–223. doi: 10.1038/hdy.2015.91

Factor de impacto: 3.801 (2015)

Posición: 29/149 (Q1) *Ecology*

45/166 (Q2) *Genetics & Heredity*

12/45 (Q2) *Evolutionary Biology*

2

Calvo-Martín JM, Papaceit M, Segarra C (2017) Evidence of neofunctionalization after the duplication of the highly conserved Polycomb group gene *Caf1-55* in the obscura group of *Drosophila*. *Scientific Reports* 7:40536. doi: 10.1038/srep40536

Factor de impacto: 5.228 (2015)

Posición: 7/63 (Q1) *Multidisciplinary Sciences*

3

Calvo-Martín JM, Papaceit M, Segarra C Molecular population genetics of the Polycomb genes in *Drosophila subobscura*.

Pendiente de publicación

4

Calvo-Martín JM, Papaceit M, Segarra C (2017) *Sex combs reduced* (Scr) regulatory region of *Drosophila* revisited. *Molecular Genetics and Genomics*. doi: 10.1007/s00438-017-1309-1

Factor de impacto: 2.622 (2015)

Posición: 75/166 (Q2) *Genetics & Heredity*

147/ 289 (Q3) *Biochemistry & Molecular Biology*

En el artículo 1 el doctorando contribuyó al diseño del estudio, realizó todo el trabajo experimental, llevó a cabo la mayor parte del trabajo analítico a excepción de la labor computacional en el análisis coevolutivo, y participó en la escritura del manuscrito. En los artículos 2, 3 y 4 el doctorando contribuyó al diseño de cada estudio, realizó la totalidad del trabajo experimental y analítico, y participó en la escritura de los respectivos manuscritos. En ningún caso se han utilizado, implícita o explícitamente, los artículos presentados en esta tesis para la elaboración de otras tesis doctorales.

Dra. Carmen Segarra

Profesora Titular
Departamento de Genética,
Microbiología y Estadística
Facultad de Biología
Universidad de Barcelona

Dra. Montserrat Papaceit

Profesora Agregada
Departamento de Genética,
Microbiología y Estadística
Facultad de Biología
Universidad de Barcelona

Artículos

Artículo 1

Adaptive selection and coevolution at the proteins of the Polycomb repressive complexes in *Drosophila*

Calvo-Martín JM, Librado P, Aguadé M, Papaceit M, Segarra C

2016. *Heredity* 116:213–223. doi: 10.1038/hdy.2015.91

Las proteínas del grupo Polycomb (PcG) constituyen un sistema clave de regulación epigenética, modulando el estado de la cromatina mediante modificaciones postraduccionales de las histonas. Las proteínas Polycomb reprimen la expresión génica actuando en forma de complejos multiméricos, por lo que se espera que evolucionen de forma coordinada. Esta coevolución podría verse reflejada en sus árboles filogenéticos a modo de episodios concordantes de selección adaptativa, así como por una correlación de sus tasas evolutivas. Con el objetivo de detectar estas señales de coevolución se ha analizado la evolución molecular de 17 genes que codifican las subunidades de cinco complejos represores Polycomb en el género *Drosophila*. La distribución observada de su divergencia difiere substancialmente entre proteínas y a lo largo de ellas. Así, CAF1 (CAF1-55) está uniformemente conservada, mientras que en proteínas como PHO, PHOL, PSC, PH-P y ASX únicamente aparecen conservados los dominios proteicos previamente descritos. Además, en SFMBT y SU(Z)12 se han detectado regiones con una reducida divergencia pese a no estar descritas previamente como dominios proteicos. Los métodos de máxima verosimilitud empleados indican una aceleración en la tasa de sustitución no sinónima en el linaje ancestral de las especies del grupo *obscura*, compartida por la mayor parte de los genes que codifican para las subunidades del complejo Pcl-PRC2, así como por los genes *Sfmbt*, *Psc* y *Kdm2*. Estos métodos también permiten inferir episodios puntuales de selección adaptativa en los genes *E(z)* y *Sfmbt* en el mencionado linaje. Por último, la red de interacciones inferidas a partir de los proteomas completos de 12 especies de *Drosophila*, usando una aproximación coevolutiva, muestra dos agrupaciones de genes Polycomb bien definidas. Ambos grupos incluyen tanto interacciones binarias previamente identificadas como nuevas interacciones que todavía no han sido confirmadas experimentalmente.

ORIGINAL ARTICLE

Adaptive selection and coevolution at the proteins of the Polycomb repressive complexes in *Drosophila*

JM Calvo-Martín, P Librado, M Aguadé, M Papaceit and C Segarra

Polycomb group (PcG) proteins are important epigenetic regulatory proteins that modulate the chromatin state through posttranslational histone modifications. These interacting proteins form multimeric complexes that repress gene expression. Thus, PcG proteins are expected to evolve coordinately, which might be reflected in their phylogenetic trees by concordant episodes of positive selection and by a correlation in evolutionary rates. In order to detect these signals of coevolution, the molecular evolution of 17 genes encoding the subunits of five Polycomb repressive complexes has been analyzed in the *Drosophila* genus. The observed distribution of divergence differs substantially among and along proteins. Indeed, CAF1 is uniformly conserved, whereas only the established protein domains are conserved in other proteins, such as PHO, PHOL, PSC, PH-P and ASX. Moreover, regions with a low divergence not yet described as protein domains are present, for instance, in SFMBT and SU(Z)12. Maximum likelihood methods indicate an acceleration in the nonsynonymous substitution rate at the lineage ancestral to the obscure group species in most genes encoding subunits of the Pcl-PRC2 complex and in genes *Sfmbt*, *Psc* and *Kdm2*. These methods also allow inferring the action of positive selection in this lineage at genes *E(z)* and *Sfmbt*. Finally, the protein interaction network predicted from the complete proteomes of 12 *Drosophila* species using a coevolutionary approach shows two tight PcG clusters. These clusters include well-established binary interactions among PcG proteins as well as new putative interactions.

Heredity (2016) 116, 213–223; doi:10.1038/hdy.2015.91; published online 21 October 2015

INTRODUCTION

Polycomb group (PcG) proteins constitute an epigenetic silencing system with a key role in the stable transcriptional repression of homeotic (*Hox*) genes, X chromosome inactivation, genomic imprinting, prevention of senescence, stem cell regulation, cell fate determination and cancer (as reviewed in Simon and Kingston, 2013). Although first identified in *Drosophila melanogaster*, these proteins are present from fungi to mammals and plants. PcG proteins have been purified as subunits of diverse multimeric complexes that modulate the chromatin state around the target genomic regions called Polycomb response elements (PREs) through posttranslational histone modifications. Genes silenced by the PcG complexes on their binding to PREs are mainly transcription factors and signaling pathway components.

In *Drosophila*, the two main Polycomb repressive complexes (PRCs) are PRC2 and PRC1. The core of PRC2 contains four proteins: enhancer of zeste (*E(Z)*), extra sex combs (*ESC*) or alternatively *EscL* (*ESCL*), suppressor of zeste 12 (*SU(Z)12*) and chromatin assembly factor 1 subunit (*CAF1*, also named *NURF55*). *E(Z)* has a SET domain and catalyzes the mono-, di- and trimethylation of histone H3 at lysine 27, which is the repressive chromatin hallmark. *ESC* boosts the enzymatic activity of *E(Z)*, whereas *SU(Z)12* and *CAF1* are essential for nucleosome binding. A distinct form of the PRC2 complex contains an additional protein: polycomb-like (*PCL*). This five-protein complex (*Pcl-PRC2*) is responsible for the high levels of H3K27 trimethylation (*H3K27me3*) around the PREs that are needed to maintain the repressed state (Nekrasov *et al.*, 2007).

The core of PRC1 also contains four proteins: sex combs extra (*SCE*, also named *dRING*), posterior sex combs (*PSC*), polycomb (*PC*) and polyhomeotic (*PH*). *SCE* has an E3 ubiquitin ligase activity and it monoubiquitylates lysine 118 of histone H2A (*H2AK118ub*), which is a second chromatin silencing mark. *PC* has a chromodomain that specifically recognizes the H3K27me3 hallmark deposited by *Pcl-PRC2*, whereas *PSC* and *PH* inhibit chromatin remodeling. PRC1 core complexes can be copurified with the sex comb on midleg (*SCM*) protein that could contribute to the recruitment of PRC1 at PRE (Wang *et al.*, 2010).

The *SCE* and *PSC* proteins of PRC1 are two components of another Polycomb complex named *dRAF* (*dRING*-associated factors) that also contains the lysine (K)-specific demethylase 2 (*KDM2*) protein (Lagarou *et al.*, 2008). This last protein enhances the ubiquitin ligase activity of *SCE* and has a *JmjC* demethylase domain, which mediates the demethylation of H3K36 that in its methylated state is a signal of gene activation. Therefore, PcG proteins do not only incorporate silencing hallmarks in the chromatin but they also remove activation marks introduced by proteins of the Trithorax group. Repression of PcG target genes also requires the activity of the *PR-DUB* complex (Scheuermann *et al.*, 2010) that includes the calypso and additional sex combs (*ASX*) proteins and has an H2A deubiquitinase activity. Thus, surprisingly both H2A monoubiquitination and deubiquitination have a role in Polycomb repression.

PhoRC (*Pho* repressive complex) is an additional Polycomb complex with two components: pleiohomeotic (*PHO*), or alternatively

Facultat de Biologia, Departament de Genètica, and Institut de Recerca de la Biodiversitat (IRBio), Universitat de Barcelona, Barcelona, Spain
 Correspondence: Dr C Segarra, Facultat de Biologia, Departament de Genètica, Universitat de Barcelona, Diagonal 643, Barcelona 08028, Spain.
 E-mail: csegarra@ub.edu

Received 22 January 2015; revised 23 July 2015; accepted 10 August 2015; published online 21 October 2015

Pho-like (PHOL), and the product of the *Scm-related gene containing four MBT domains* gene (SFMBT). PHO and PHOL are the only two described PcG proteins with a DNA-binding affinity and thus PhoRC is crucial for anchoring the other PRCs at PREs. However, the process of recruitment of PcG complexes to chromatin is not fully characterized. According to the classical model, PhoRC would have a pivotal role in the initial recruitment of PRC2 to PREs. Thereafter, the H3K27me3 hallmark deposited by Pcl-PRC2 would trigger the recruitment of PRC1 by means of the chromodomain of PC. Nevertheless, PHO interacts not only with subunits of the PRC2 complex (Wang *et al.*, 2006) but also with subunits of the PRC1 complex (Mohd-Sarip *et al.*, 2002), which could also contribute to anchor PRC1 at PRE. In addition, it has been proposed that the recruitment of PcG complexes to chromatin might require additional not yet identified proteins (Wang *et al.*, 2010). The genome-wide identification of PRE regions by chromatin immunoprecipitation assays in *Drosophila* identified ~200 large Polycomb domains where PcG proteins colocalize (Schuettengruber *et al.*, 2009, and references therein). These elements are included in broader chromatin regions with the H3K27me3 hallmark. The characterized PREs do not share any sequence similarity among them, but contain multiple binding motifs for different DNA-binding proteins (including PHO) that would work combinatorially to recruit PcG complexes (Schuettengruber and Cavalli, 2009). Interestingly, PcG proteins bound to a PRE can silence rather distant promoters, which indicates that chromatin looping is important in the regulation by PcG proteins. In addition, PcG proteins are concentrated in nuclear foci termed PcG bodies, suggesting long-range chromosomal interactions to form distinct repressive compartments in the interphasic nuclei (Pirrotta and Li, 2012).

The multiple interactions among PcG proteins required to form the repressive complexes and to act in concert to repress chromatin might prevent their independent evolution. In fact, interacting proteins in protein-protein networks tend to evolve at similar rates (reviewed by Lovell and Robertson, 2010). Species with a well-characterized interactome, such as yeast, have provided the strongest support for this trend, given that interacting proteins exhibit highly correlated evolutionary rates (Hakes *et al.*, 2007). The correlation of evolutionary rates can be the result of molecular coevolution (or coadaptation according to Juan *et al.*, 2008) between the residues of interacting proteins. Indeed, the substitution of an amino acid in the interaction interface of one protein might be compensated in the binding partner protein by another substitution that maintains the integrity and functionality of the protein complex.

A wide range of bioinformatic methods have been designed to detect coevolution at the molecular level and thus to predict protein interactions (reviewed in Juan *et al.*, 2013). Mirror Tree was the first method developed to analyze the correlation of evolutionary rates between whole proteins. Initially, this method quantified the extent of the interaction between two proteins through the correlation coefficient of the distance matrix of each protein as inferred from the multiple alignments of the orthologous sequences in the same set of species. This method was later modified to correct by the background similarity expected even between non-interacting proteins due to phylogenetic relationships, that is, the common ancestry of the species included in the data set (Fraser *et al.*, 2004; Clark and Aquadro, 2010). An additional modification that also improved Mirror Tree performance is the Context Mirror approach that takes into account the possibility of multiple protein interactions and allows the detection of only those interactions specific to a single protein pair also in an evolutionary context (Juan *et al.*, 2008).

Here we analyze the molecular evolution of 17 genes encoding proteins of the Polycomb complexes PhoRC, Pcl-PRC2, PRC1, dRAF and PR-DUB in the 12 *Drosophila* species with whole sequenced genomes (Clark *et al.*, 2007) and in 3 additional species of the obscura group (*Drosophila subobscura*, *Drosophila madeirensis* and *Drosophila guanche*). The main aim of this study is to analyze whether gene divergence in the *Drosophila* genus is affected or not by the physical interactions of the encoded proteins either to form the complexes or to assist in their coordinated function. Indeed, the proteins that form these complexes are good candidates to detect correlated molecular evolution. The results obtained indicate that: (i) the distribution of divergence differs substantially among and across PcG proteins, CAF1 being the most uniformly conserved protein; (ii) highly conserved regions not yet described as protein domains but that might be relevant to protein function are present in SFMBT and SU(Z)12; (iii) an increase in the nonsynonymous divergence in the lineage ancestral to the obscura group species has driven the molecular evolution of most genes coding for subunits of the Pcl-PRC2 complex and also of genes *Sfmbt*, *Psc* and *Kdm2*; (iv) positive selection has acted in this lineage at genes *E(z)* and *Sfmbt*; and (v) there are two tight clusters of coevolving PcG proteins: one with ASX and the subunits of the Pcl-PRC2 complex and another with CALYPSO and subunits of the PhoRC, PRC1 and dRAF complexes.

MATERIALS AND METHODS

Fly samples and DNA sequencing

The *ch cu* strain of *D. subobscura* and highly inbred lines of *D. madeirensis* and *D. guanche* established after 10 generations of sib mating were available in our laboratory. DNA from these lines was purified with the PuregenCore Kit B (Qiagen, Hilden, Germany) according to the manufacturer's instructions. The 17 PcG genes from each species were PCR amplified with primers designed with the OLIGO v4.1 program (Rychlik, 1993) on the *Drosophila pseudoobscura* sequence available at flybase (www.flybase.org). Amplicons were purified with Qiaquick columns (Qiagen) and sequenced using the ABI Prism BigDye Terminators 3.0 Cycle sequencing kit (Applied Biosystems, Waltham, MA, USA). Sequencing reactions were run on an ABI PRISM 3700 automated DNA sequencer. Both strands were completely sequenced by primer walking using internal primers. The sequences of the primers used in the PCR amplification and sequencing, as well as the PCR conditions, are available on request to the authors. The Seqman program of the DNASTAR Lasergene package (Burland, 2000) was used to assemble the partial sequences.

Divergence analysis

The sequences of the 17 Polycomb genes in the 12 *Drosophila* species with sequenced genomes (Clark *et al.*, 2007) were retrieved from flybase (www.flybase.com) after Blast searches. The MUSCLE program (Edgar, 2004) implemented in the software MEGA6 (Tamura *et al.*, 2013) was used to multiple align the orthologous sequences of each gene according to the alignment of the predicted proteins. Alignments were manually checked and corrected whenever necessary. The identification of domains in the *D. melanogaster* protein sequences was performed using the InterProScan v5 program (Jones *et al.*, 2014) available in the EMBL-EBI web site (<http://www.ebi.ac.uk/Tools/pfa/iprscan5>). This program combines different protein signature databases and recognition methods into one resource. Additional protein motifs relevant for protein function described in the literature but not included in these databases were also identified. The putative presence of intrinsically disordered regions across the *D. melanogaster* proteins was analyzed with the PONDR-FIT predictor (Xue *et al.*, 2010) available at <http://www.disprot.org>.

The Clustal X v2.1 program (Larkin *et al.*, 2007) was used to infer the scores of amino acid conservation at each site of the multiple alignment of a given protein according to the protein weight matrix BLOSUM62. This analysis was performed with the raw protein alignments including regions with indels and with an uncertain alignment. These regions were manually excluded from the alignment in the subsequent analyses. The accepted phylogenetic tree of the

studied species was analyzed by maximum likelihood with the MEGA6 program (Tamura et al., 2013), to infer and visualize the branch lengths of each protein phylogeny, according to the Jones–Taylor–Thornton model.

Selection analysis

The PAML v4 package (Yang, 2007) was used to compare different evolutionary models with alternative assumptions on the ω value in the accepted phylogeny of the studied species. ω estimates ($\omega = d_N/d_S$, where d_N and d_S correspond to nonsynonymous and synonymous divergence, respectively) were inferred with the CODEML program implemented in PAML. The M0 model that assumes the same ω value in all branches was compared with the free ratio model (FR) that considered a different ω in each branch. The M0 model was also compared with the branch model 2R that assumes that the ω value for particular branches of the phylogeny defined as foreground branches differs from the ω value of the rest of branches (background branches). The branch-site test of positive selection or test 2, as defined by Zhang et al. (2005), was also applied to detect the presence of positively selected sites in the foreground branch. In this test, the modified branch-site model A that includes a site class with $\omega_2 > 1$ (that is, under positive selection) in the foreground branch is compared with a null model with a fixed $\omega_2 = 1$ in this branch. For each model, the CODEML program implemented in PAML was run multiple times with different initial values to prevent incorrect parameter estimates caused by local optima. In all cases, a likelihood ratio test (LRT) was used to infer whether the null model could be rejected assuming that twice the log likelihood difference between two nested models differing in n free parameters follows a χ^2 -distribution with n degrees of freedom. Sites under positive selection in the foreground branch were identified by the Bayes Empirical Bayes method (Yang et al., 2005). The random effects branch-site model (REL) developed by Kosakovsky Pond et al. (2011) and available in the datamonkey web server (www.Datamonkey.org) was also applied to each gene, to detect branch-specific signals of episodic selection in the phylogeny.

Coevolution analysis

The Context Mirror approach (Juan et al., 2008), based on the Mirror Tree method, was used to detect putative coevolution among the studied Polycomb proteins. With this purpose, the complete sets of protein-coding genes for the 12 *Drosophila* species, as well as their pairwise orthologous relationships, were downloaded from FlyBase (*D. melanogaster* release 5.56, *Drosophila simulans* r1.4, *Drosophila sechellia* r1.3, *Drosophila yakuba* r1.3, *Drosophila erecta* r1.3, *Drosophila ananassae* r1.3, *Drosophila persimilis* r1.3, *D. pseudoobscura* r3.1, *Drosophila willistoni* r1.3, *Drosophila grimshawi* r1.3, *Drosophila mojavensis* r1.3, and *Drosophila virilis* r1.2). The retrieved sequences were replaced when required by those reannotated or resequenced in the present study. Pairs of homologous genes were clustered into groups of multiple species with the mcl program available at <http://micans.org/mcl>. Only orthologous groups with a single gene copy (that is, 1:1) in each *Drosophila* species were retained for further analyses ($n = 6563$). For each 1:1 group, the most likely orthologous combination of protein isoforms was selected using the PALO software (Villanueva-Cañas et al., 2013). These protein isoforms were aligned using the probabilistic framework provided by PRANK v1.4 (Löytynoja and Goldman, 2008). Perl scripts were developed to filter out alignment positions with a posterior probability lower than 99%. The amino acid distances among *Drosophila* species were estimated with the CODEML program of the PAML v4 package (Yang, 2007), separately for each orthologous group. As a proxy for coevolution, these amino acid substitution rates were compared across phylogenetic lineages, using the ContextMirror program with P -value cutoffs of 10^{-4} (Juan et al., 2008). Cytoscape (Shannon et al., 2003) was used to represent significant instances of coevolution as a network.

RESULTS

Gene identification

Molecular evolution of 17 genes encoding subunits of five PRCs has been analyzed. The studied PcG genes are: *Sfmbt*, *Pho* and its paralog *Phol* (PhoRC complex); *E(z)*, *Esc*, its paralog *Escl*, *Su(z)12*, *Caf1* and *Pcl* (Pcl-PRC2 complex); *Scs*, *Psc*, *Pc*, *Ph-p* and *Scm* (PRC1 complex);

Kdm2 (dRAF complex); and *Calypso* and *Asx* (PR-DUB complex). These genes were identified in *D. melanogaster*, *D. simulans*, *D. sechellia*, *D. yakuba*, *D. erecta*, *D. ananassae*, *D. pseudoobscura*, *D. persimilis*, *D. willistoni*, *D. mojavensis*, *D. virilis* and *D. grimshawi* after Blast searches against their complete genomes made available by the *Drosophila* 12 Genomes Consortium Initiative (Clark et al., 2007). The first nine species are members of the *Sophophora* subgenus, whereas the last three belong to the *Drosophila* subgenus. The 17 genes were present in the 12 species. The paralog of *Ph-p* (*Ph-d*) was not included in the analysis given its absence in the species of the *Drosophila* subgenus. Several cases of sequencing errors and misspellings were detected (Supplementary Table S1). Most errors are single nucleotide indels that cause frameshift mutations and occasionally affect the predictions of the gene intron–exon boundaries. In addition, there are some cases where the sequence of a gene was partially incomplete in a species. Almost all these problems could be corrected by resequencing the corresponding gene region. The complete coding region of the 17 genes was also sequenced in the three closely related species *D. subobscura*, *D. madeirensis* and *D. guanche* (*subobscura* subgroup) that like *D. pseudoobscura* and *D. persimilis* are included in the *obscura* group of *Drosophila*.

The exon–intron organization of most genes (*Phol*, *Caf1*, *E(z)*, *Su(z)12*, *Escl*, *Psc*, *Scs*, *Pc*, *Scm* and *Calypso*) as annotated in *D. melanogaster* is conserved in the studied species (Figure 1). In contrast, intron gain or loss events were detected in particular lineages at *Sfmbt*, *Esc*, *Pcl*, *Ph-p*, *Kdm2* and *Asx*. Finally, in *Pho* the exact boundaries of exons 2 and 3 could not be confirmed with certainty in all species given the high divergence of the gene in this region. Each gene is included in scaffolds assigned to the same Muller's element in the 12 species with complete sequenced genomes.

Divergence along the subunits of the PcG complexes

The distribution of the amino acid divergence along the multiple alignment of each protein is shown in Figure 2, where the described protein domains are also indicated. The PhoRC complex is formed by two subunits: PHO (or PHOL) and SFMBT. Divergence at PHO is rather high with some multiple alignment positions being quite uncertain. Only the four C2H2-type zinc fingers and the SP (spacer) domains (Lesley Brown et al., 1998) are highly conserved. SP is a functionally characterized domain that is critical for the interaction of PHO with E(Z) and participates in the interaction with ESC (Wang et al., 2004). PHOL shows a pattern of divergence similar to that of PHO, although it is not so divergent among species. SFMBT has an FCS-type zinc finger domain, four MBT (malignant brain tumor) repeats and a SAM (sterile α -motif) domain that are highly conserved.

The most conserved subunit of the Pcl-PRC2 complex is CAF1 (Figure 2), a member of the WD40 family with a seven-bladed β -propeller structure. Only 7 amino acid replacements (5 of them present in a single species) were detected in the 430 residues long CAF1 protein. The ESC and ESCL proteins that, like CAF1, are members of the WD40 family are also highly conserved not only in the characteristic WD40 region but also in the NTD domain (N-terminal domain), which in ESC binds to histone H3, and thus might contribute to enhance the enzymatic activity of E(Z). The SET catalytic domain of E(Z) characteristic of histone methyltransferases is extremely conserved. Indeed, no replacements were detected in this domain in the 15 species studied. The other E(Z) domains (O'Connell et al., 2001) also show a low divergence (that is, the CXC domain preceding SET, EID (ESC interacting domain), PBD1, PBD2 (PCL binding domains), SANT and domain II that interacts with SU(Z)12). The three domains of SU(Z)12 are also highly conserved: the zinc



Figure 1 Comparative organization of the coding region in the studied Polycomb genes grouped by Polycomb complex in the 15 *Drosophila* species. Conserved exons are represented by black boxes that indicate the direction of transcription and conserved introns by lines. Empty white boxes indicate nonconserved exons and white boxes with internal lines indicate nonconserved introns. The length of the conserved introns interrupted by dashes is not proportional to the scale.

finger C2H2, VEFs box and the NTD that interacts with CAF1 (Nowak *et al.*, 2011). The presence in SU(Z)12 of two highly conserved regions between NTD and C2H2 not yet characterized as protein domains is noteworthy. PCL is the most divergent subunit of the Pcl-PRC2 complex. Protein conservation is nearly restricted to the described protein domains: Tudor, the two PHD-type zinc fingers, CLD (chromo-like domain; Wang *et al.*, 2004) and EH domain (extended homology domain; Wang *et al.*, 2004).

The core of the PRC1 complex has four subunits: PSC, SCE, PC and PH-P. Divergence along PSC and PH-P is distributed quite heterogeneously and highly divergent regions are interspersed with conserved regions, some of which correspond to the described protein domains: the RING-type zinc finger and the HTH (helix-turn-helix domain; Kyba and Brock, 1998) of PSC, and the FCS zinc finger and SAM domains of PH-P. Conservation is high around the RING-type zinc finger and the two HD domains (Gorfinkel *et al.*, 2004) of SCE, which, in turn, has a rather high divergence outside the functional domains. A similar pattern is detected in SCM, where a high conservation is detected mainly in the protein domains (that is, the FCS zinc finger, the two MBT repeats, the Scm-like embedded domain (SLED) and the SAM domains). The chromodomain (CD) of PC that recognizes the H3K27me3 hallmark established by the E(Z) subunit of PRC2 is highly conserved, as it is its C-terminal domain (CTD; Franke *et al.*, 1995) and also a central region not previously identified as a protein domain. Divergence along KDM2 (that together with PSC and SCE forms the dRAF complex) shows highly conserved regions that include the JmjC catalytic domain, the CXXC-type zinc finger,

the F-box and the Amn1/LRR domains. Finally, the pattern of divergence differs substantially along the two subunits of the PR-DUB complex. CALYPSO with a peptidase C12-ubiquitin carboxyl-terminal hydrolase domain is highly conserved, whereas conservation along ASX is almost only restricted to the ASXH and the PHD-type zinc finger domains.

Selection in the subunits of the PcG complexes

Protein divergence may vary in different lineages due to episodic adaptive selection resulting in an acceleration of the amino acid substitution rate, which can be reflected in the protein phylogenetic tree. In addition, in the case of interacting proteins the same lineages might be affected by positive selection due to protein coevolution. As a first approach to detect adaptive selection, phylogenetic trees were obtained for each PcG protein by maximum likelihood with the MEGA6 program (Tamura *et al.*, 2013) fixing the commonly accepted phylogeny of the 15 studied species, in order to infer branches lengths and therefore the substitution rate for each branch. The most striking result of this analysis is the strong acceleration in the amino acid substitution rate at the branch ancestral to the species of the obscura group (henceforth, *Bobs* branch or lineage) for three proteins of the Pcl-PRC2 complex: the two interacting proteins E(Z) and ESC, and PCL that interacts with E(Z). The same tendency, although not so strong, is detected in SFMBT of the PhoRC complex (Figure 3). Phylogenetic trees based on synonymous divergence in the corresponding genes do not exhibit this acceleration (results not shown), which indicates a decoupling between the synonymous and

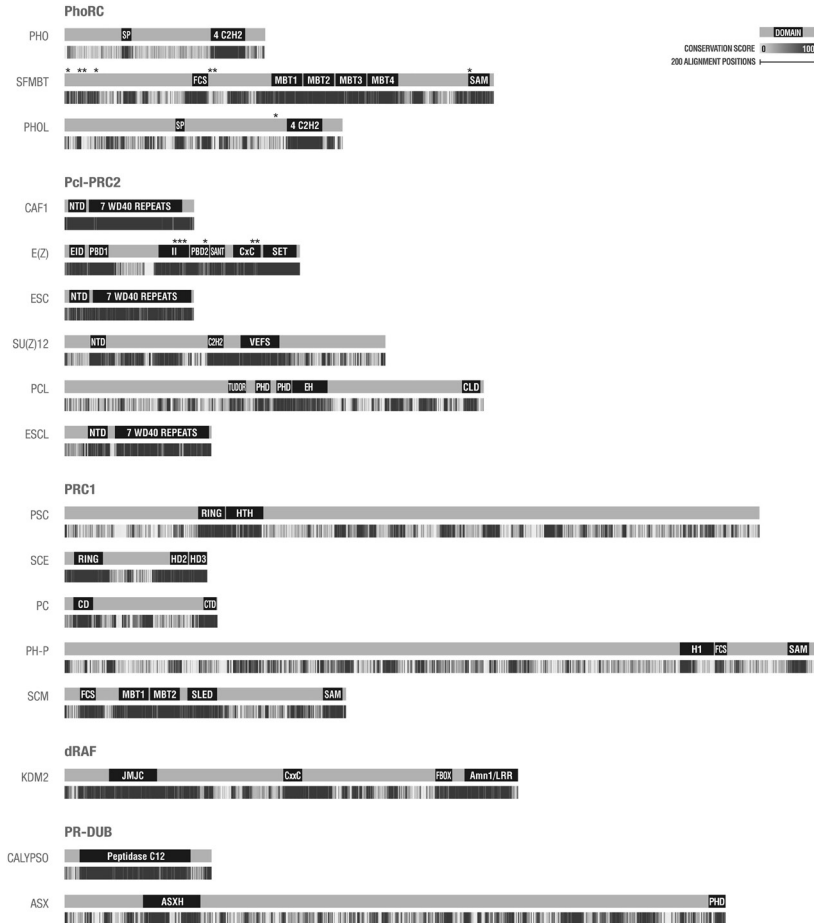


Figure 2 Protein conservation plots inferred with Clustal X v2.1 (Larkin *et al.*, 2007) along the multiple alignment of the studied Polycomb proteins grouped by Polycomb complex. Gray intensity indicates the conservation score of each protein position from 0 (white) to 100 (black). The bar above each conservation plot shows the described protein domains. Asterisks above the bar indicate amino acid positions under positive selection inferred by the Bayes Empirical Bayes method implemented in PAML.

nonsynonymous substitution rate in the *Bobs* branch. This result suggests positive selection acting on these proteins before the split of the *subobscura* and *pseudoobscura* subgroups and it even suggests the coevolution of some PcG proteins.

According to these results, the evolutionary models implemented in PAML (Yang, 2007) were applied to try to detect adaptive selection acting on PcG genes in particular lineages of the phylogeny. As a first approach, the M0 model, which assumes the same ω estimate ($\omega = d_N/d_S$, that is, the ratio between nonsynonymous and synonymous divergence) for all branches was compared with the free ratio (FR) model, which allows a different ω for each branch. Except for genes *Caf1*, *Pc*, *Scm* and *Calypso*, the M0 model could be rejected

(Table 1), indicating heterogeneity in the ω estimates across branches in the phylogeny for most studied genes. Next, the M0 model was compared with a branch model ($2R-f_{Bobs}$) assuming two ω estimates: one for the branch leading to the *obscura* group species (*Bobs*) that was fixed as foreground branch ($\omega_{f_{Bobs}}$) and one for the rest of branches (ω_b). The $2R-f_{Bobs}$ model is better supported by the data than the M0 model, not only for the four genes with a long *Bobs* branch in Figure 3 (*E(z)*, *Esc*, *Pcl* and *Sfmbt*) but also for *Su(z)12*, *Psc*, *Sce* and *Kdm2*. In *Esc* and *Su(z)12*, ω heterogeneity is exclusively explained by the *Bobs* branch, as the $2R-f_{Bobs}$ model could not be rejected when compared with the free ratio (FR) model (Table 1). The ML estimates under the best-fit model reveal that $\omega_{f_{Bobs}}$ values are

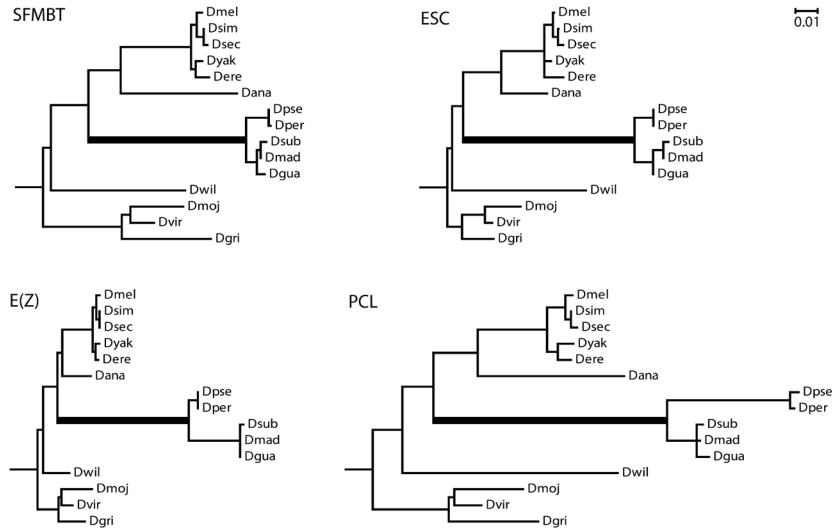


Figure 3 Phylogenetic trees according to protein divergence of the SFMBT subunit of the PhoRC complex and of the E(Z), ESC and PCL subunits of the Pcl-PRC2 complex. Branch lengths were obtained with MEGA6 (Tamura *et al.*, 2013) by maximum likelihood fixing the commonly accepted phylogeny of the studied species. The branch leading to the *Bobs* group species (*Bobs* in the text) is highlighted. Dmel, *D. melanogaster*; Dsim, *D. simulans*; Dsec, *D. sechellia*; Dyak, *D. yakuba*; Dere, *D. erecta*; Dana, *D. ananassae*; Dpse, *D. pseudoobscura*; Dper, *D. persimilis*; Dsub, *D. subobscura*; Dmad, *D. madeirensis*; Dgua, *D. guanche*; Dwil, *D. willistoni*; Dmoj, *D. mojavensis*; Dvir, *D. virilis*; and Dgri, *D. grimshawi*.

significantly higher than *ob* values in four of the six genes coding for subunits of the Pcl-PRC2 and also in *Sfmbt*, *Psc* and *Kdm2* (Figure 4). The highest difference is detected in *Esc* and *E(z)* with a $\omega_{\text{Bobs}}/\omega_{\text{ob}}$ ratio equal to 13.9 and 8.7, respectively. Finally, the branch site test of positive selection (or test 2 in Zhang *et al.*, 2005) was applied, which showed that the differences at *E(z)*, *Sfmbt* and *Phol* in the ω_{Bobs} and *ob* estimates are at least partly explained by positive selection acting in the *Bobs* lineage. In addition, the Bayes empirical Bayes method identified six codons with a high posterior probability of having evolved under positive selection in the *Bobs* lineage at *E(z)*, seven codons at *Sfmbt* and one at *Phol* (Figure 2). All sites under positive selection at *E(z)* are located in gene regions that code for described protein domains. In contrast, only one of the selected codons at *Sfmbt* is in a region coding for a domain, but two of them are close to one such region. The remaining four codons cluster in a region at the beginning of the open reading frame and thus at the conserved N-terminal region of the protein. The single codon under positive selection at *Phol* is neither located in a conserved region nor in a region coding for a protein domain.

The described PAML analysis was performed fixing as foreground the *Bobs* branch suspected to have been under positive selection at least in some genes according to the phylogenetic trees (Figure 3). Kosakovsky Pond *et al.* (2011) developed a random-effects branch site method to detect episodic positive selection in particular lineages of a phylogeny that does not require defining *a priori* foreground and background branches. This likelihood approach (REL) was applied to each gene, to further corroborate the positive selection results. *E(z)*, *Esc*, *Phol*, *Sfmbt* and *Ph-p* showed evidence of episodic selection in the *Bobs* lineage (*P*-value = 0.001, 0.010, 0.042, 0.001 and 0.015, respectively). However, *P*-values remain significant only for *E(z)* and *Sfmbt*

when correcting for multiple testing (corrected *P*-value = 0.014 and 0.037, respectively). Therefore, both PAML and REL maximum likelihood approaches infer the action of positive selection in the *Bobs* lineage in *E(z)* and *Sfmbt*.

Coevolution of the PcG proteins

The proteomes of the 12 *Drosophila* species with complete sequenced genomes (Clark *et al.*, 2007) were analyzed for the first time using the Context Mirror approach (Juan *et al.*, 2008), to uncover any evidence of coevolution among the PcG proteins. The interaction network of the proteins with significant partial correlations of evolutionary rates and thus predicted to coevolve is shown in Figure 5. This network includes only 16 of the 17 PcG proteins here studied, as PHOL did not reach the established *P*-value cutoff of significance. PcG proteins form two tight clusters of coevolving proteins (Figure 5). One predicted cluster interconnects linearly four of the five subunits of the Pcl-PRC2 complex: E(Z), ESC, PCL and SU(Z)12. There is, therefore, evidence that these four proteins have coevolved. The other two proteins of this complex (CAF1 and ESCL) are not included in this cluster. Interestingly, the Pcl-PRC2 cluster also includes ASX, one of the subunits of PR-DUB. The second predicted cluster interconnects CALPYSO (the other subunit of PR-DUB) and all the proteins of the PhoRC, PRC1 and dRAF complexes. Thus, PHO and SFMBT (PhoRC complex) are predicted to interact not only between them but also with the five proteins of the PRC1 complex (PC, PSC, SCE, PH-P and SCM), which, in turn, are predicted to interact among them. In addition, CALPYSO and the three proteins of dRAF (KDM2, SCE and PSC) are also interconnected both among them and with the proteins of PRC1. These results clearly indicate the power of the

Table 1 *P*-values of the LRTs to contrast different evolutionary models implemented in PAML (Yang, 2007)

PcG complex gene	Models ^a			
	MO vs FR	MO vs 2R- <i>f</i> _{Bobs}	2R- <i>f</i> _{Bobs} vs FR	Test 2
<i>PhoRC</i>				
<i>Pho</i>	0.016*	0.449	0.013*	0.500
<i>Sfmbt</i>	0.000***	0.000***	0.044*	0.033*
<i>Phol</i>	0.002***	0.068	0.004***	0.017*
<i>Pcl-PRC2</i>				
<i>Caf1</i>	0.201	0.417	0.186	0.500
<i>E(z)</i>	0.000***	0.000***	0.001***	0.031*
<i>Esc</i>	0.000***	0.000***	0.439	0.231
<i>Su(z)12</i>	0.040*	0.004***	0.205	0.409
<i>Pcl</i>	0.000***	0.000***	0.000***	0.500
<i>Esc1</i>	0.004***	0.309	0.004***	0.175
<i>PRC1</i>				
<i>Psc</i>	0.000***	0.014*	0.000***	0.500
<i>Sce</i>	0.000***	0.013*	0.001***	0.500
<i>Pc</i>	0.742	0.376	0.735	0.072
<i>Ph-p</i>	0.001***	0.427	0.001***	0.063
<i>Scm</i>	0.056	0.315	0.053	0.500
<i>dRAF^b</i>				
<i>Kdm2</i>	0.000***	0.006**	0.000***	0.215
<i>PR-DUB</i>				
<i>Calypso</i>	0.780	0.134	0.859	0.500
<i>Asx</i>	0.000***	0.112	0.000***	0.500

Significance: *0.05 > *P* > 0.01; **0.01 > *P* > 0.005; ****P* < 0.005.

^aMO, a single ω for all branches; free ratio (FR), a different ω for each branch; 2R- f_{Bobs} , one ω for the branch ancestral to the obscura group species and one ω for the rest of branches. Test 2 of positive selection as described in Zhang et al. (2005).

^bdRAF complex also contains PSC and SCE.

Context Mirror approach to detect interacting proteins from the signals of coevolution reflected in their interspecific divergence.

Despite the good performance of Context Mirror in predicting the interactions between the PcG proteins, the method also displays a certain level of background noise, as it predicts some dubious interactions. Hence, the detected Pcl-PRC2 and the PhoRC/PRC1/dRAF clusters include additional proteins that have, in most cases, an uncertain Polycomb-related function. One exception is SCR (sex combs reduced) that is interconnected with CALYPSO and most proteins of the PhoRC/PRC1/dRAF cluster. It is well established (Gindhart and Kaufman, 1995) that *Scr* is a gene directly regulated by the Polycomb complexes. Our results would suggest a putative interaction of the homeotic protein SCR with Polycomb proteins to regulate some of its target genes.

DISCUSSION

Divergence among the *Drosophila* species differs substantially in the 17 Polycomb proteins studied here. The high conservation of CAF1 stands in contrast with the presence at PHO, PSC, PH-P or ASX of highly divergent regions interspersed with rather short conserved regions. However, the described domains in the different proteins are well conserved in all species without exception (Figure 2). In some cases, conserved regions extend beyond the described domains (for example, the JmjC and the CXXC-type zinc finger domains of

KDM2 and the SAM domain of SFMBT). In addition, highly conserved protein regions not yet identified as important motifs for the protein function have been detected, for instance, in the N-terminal region of SFMBT and between the NTD and C2H2 domains of SU(Z)12 (Figure 2). Although evolutionary conservation can not be considered an unequivocal signal of functional relevance, the strong constraints acting on these regions make them good candidates to have an important and conserved role in protein function. Despite the action of strong purifying selection in these conserved regions, four codons with a high posterior probability of having evolved under positive selection were identified by the Bayes Empirical Bayes method implemented in PAML (Yang et al., 2005) in the conserved N-terminal region of SFMBT. Therefore, the detected conserved motifs might be good targets to design experimental studies trying to identify interacting interfaces between proteins or major determinants of the Polycomb complexes function as chromatin modifiers.

However, sequence conservation is not an essential requirement to maintain function, as has been shown for the PcG protein PSC. The C-terminal region of PSC is poorly conserved in sequence across insects but broadly conserved in its ability to inhibit chromatin remodeling. Indeed, it is predicted to be a structurally disordered region, in which the presence of multiple patches of high positive charge and the lack of an extended stretch of contiguous negative charge is important for protein function (Beh et al., 2012). The pattern of divergence of PH-P, PCL, PHO, PHOL and ASX with rather long, poorly conserved regions across the protein is similar to that of PSC. However, when analyzing the predicted structurally disordered regions across these proteins in *D. melanogaster*, only PH-P and ASX present a rather long stretch of ~800 and ~500 residues, respectively, with a high disorder score according to the PONDR-FIT method (Xue et al., 2010). Even though selection on disordered protein might be relaxed, positive selection acting on *Ph-p* noncoding regions has been described after the detection of a selective sweep around the first intron of this gene in a study of nucleotide variation within *D. melanogaster* (Beisswanger and Stephan, 2008).

The high conservation of the four zinc finger domains responsible for the recognition and binding of PHO at the specific target sequences present at PREs reflects the selective pressures exerted on this protein domain. It is noteworthy in this context that PREs, which are complex *cis*-regulatory DNA elements containing multiple combinations of binding sites exhibit a high sequence plasticity. Indeed, interspecific comparisons show a high variation in the number, genomic location and motif design in *Drosophila* species (Hauenschild et al., 2008). The rapid evolution of PREs might be related with the high divergence of PHO in regions outside the established functional domains.

Polycomb proteins interact among them and form functional complexes that repress gene expression. Numerous biochemical studies have focused on analyzing these interactions in *D. melanogaster*, yielding a rather well-characterized Polycomb network in this species. In addition, the regions directly involved in binary interactions have been identified for different PcG protein pairs as summarized in Figure 6. Therefore, Polycomb proteins are good candidates to contrast whether interacting proteins evolve coordinately, which might be reflected by concordant episodes of positive adaptive selection in their phylogenetic trees and by signs of coevolution. Most genes (*E(z)*, *Esc*, *Su(z)12* and *Pcl*) coding for subunits of the Pcl-PRC2 complex share a pattern of molecular evolution characterized by a ω value significantly higher in the branch leading to the obscura group species (*Bobs* branch) than in the background branches. In addition, *Sfmbt*,

Psc and *Kdm2* of the PhoRC, PRC1 and dRAF complexes, respectively, show the same pattern (Figure 4). Therefore, an acceleration of the nonsynonymous divergence in the *Bobs* branch seems to have driven the evolution of several Polycomb genes, which, in turn, suggests some kind of coevolution among the encoded proteins. However, the evolutionary models implemented in PAML and REL inferred the action of positive selection only in *E(z)* and *Sfmbt* that encode subunits

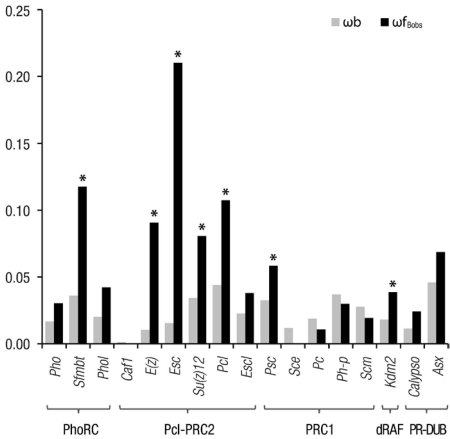


Figure 4 Maximum likelihood ω estimates inferred for each Polycomb gene by the $2R\text{-}\omega_{Bobs}$ branch model implemented in PAML v4 package (Yang, 2007) assuming two ω estimates: one fixing as foreground the branch ancestral to the obscure group species (ω_{Bobs}) and one for the rest or background branches (ω_b). Genes in which the $2R\text{-}\omega_{Bobs}$ model fits better the data than the MO model (a single ω estimate for all branches) being $\omega_{Bobs} > \omega_b$ are indicated by an asterisk. Genes are grouped according to Polycomb complex.

of different PcG complexes between which no binary direct interactions have been detected (Figure 6).

The maximum-likelihood PAML and REL approaches analyze the phylogenetic tree of each gene independently and, although concordant results in different genes might indicate a coordinated evolution, the detection of coevolution requires more powerful bioinformatic methods, such as Mirror Tree and Context Mirror. These methods directly compare the phylogenetic trees of different proteins to detect a correlated evolution due to the presence of compensatory amino acid changes required to maintain the integrity and functionality of the protein complex. It has been questioned (Hakes et al., 2007) whether the molecular coevolution of physically interacting amino acids might be reflected in the evolution of the whole protein given that the fraction of aminoacidic residues directly involved in the interaction between proteins is often small. However, several factors apart from selection on protein structure and function affect the rate of protein evolution. Some of these factors can cause similar constraints on the evolutionary rate of interacting proteins and therefore a correlated evolution. Likely, the most important of these external or indirect factors would be the similar expression level of the genes encoding interacting proteins, which ensures proper stoichiometry between the interacting components of the protein complex (Fraser et al., 2004). These factors causing similar constraints between proteins and molecular coevolution are not mutually exclusive and both can act, to some extent, to cause a correlated evolution of interacting proteins (Juan et al., 2008).

The proteomes of the 12 *Drosophila* species with complete genome sequences have been analyzed from a coevolutionary perspective focusing on the detection of the footprint of coevolution in the Polycomb proteins. The network of interacting proteins shows two tight clusters of coevolving proteins, one of them including four subunits of the Pcl-PRC2 complex (Figure 5). Among the predicted interactions in this cluster, only the binary interaction between E(Z) and ESC (E(Z)/ESC) has been well characterized experimentally (Jones et al., 1998; Tie et al., 1998). Context Mirror does not predict the proved interactions PCL/E(Z) (O'Connell et al., 2001), E(Z)/SU(Z)12

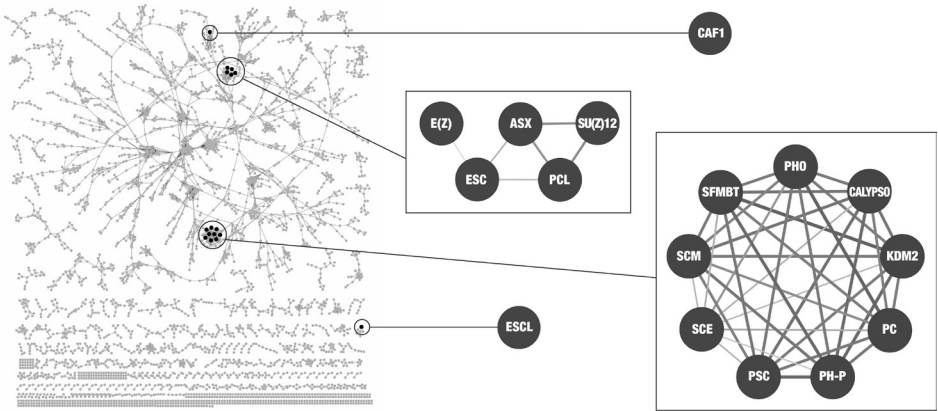


Figure 5 Left: interaction network predicted by the Context Mirror approach (Juan et al., 2008) for proteins with significant partial correlations in evolutionary rates from the 12 *Drosophila* species proteomes (Clark et al., 2007). Proteins in the network are represented by gray dots, except PcG proteins that shown by black dots. Right: zoom-in image of the predicted interactions between PcG proteins. The width and gray intensity of the connecting lines reflect the degree of the predicted interaction, which is stronger as the line gets wider and darker.

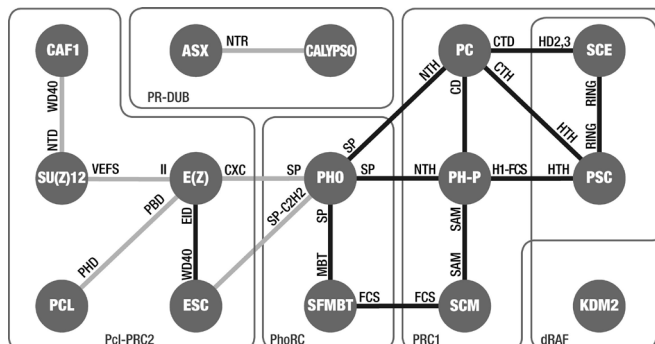


Figure 6 Binary interactions well characterized experimentally between PcG proteins of the PcI-PRC2, PhoRC, PRC1 and dRAF complexes. Interactions predicted by Context Mirror (Figure 5) are shown by black lines, in contrast to those interactions not predicted by this bioinformatic method that are shown by gray lines. The protein domains included in the regions involved in the binary interactions are shown on each line close to the corresponding protein. Protein domains are indicated as in Figure 2, except NTH (N-terminal half), CTH (C-terminal half) and NTR (N-terminal region). See text for references.

(Ketel *et al.*, 2005; Joshi *et al.*, 2008) and SU(Z)12/CAF1 (Nowak *et al.*, 2011) (Figure 6). Indeed, CAF1 is the only subunit of the Pcl-PRC2 complex that is not predicted to interact with any of the other subunits of the complex in the network, although such interactions have been experimentally inferred (Nowak *et al.*, 2011). However, CAF1 is a component of different complexes that regulate chromatin metabolism (Song *et al.*, 2008) and it is under strong functional constraints, as reflected by its high conservation. Like CAF1, ESCL is not interconnected with any other Polycomb proteins in the network. Although ESCL can be assembled *in vitro* in functional PRC2 complexes (Wang *et al.*, 2006), this result would support that *in vivo* ESCL is not a usual subunit of PRC2 when ESC is present (Kurzbaal *et al.*, 2008).

The most striking result of the coevolution analysis is that no interactions are predicted between the subunits of the PhoRC complex (PHO and SFMBT) and those of the Pcl-PRC2 complex, despite the experimental identification of the regions involved in the PHO/E(Z) and PHO/ESC interactions (Wang *et al.*, 2004). In contrast, Context Mirror predicts the PHO/SFMBT interaction (Alfieri *et al.*, 2013) and the well-characterized interactions between these two subunits of the PhoRC complex and those of PRC1: PHO/PC (Mohd-Sarip *et al.*, 2002), PHO/PH-P (Mohd-Sarip *et al.*, 2002) and SFMBT/SCM (Grimm *et al.*, 2009). The bioinformatic method also predicts the well-established interactions between subunits of the PRC1 complex, that is, PH-P/PC (Strutt and Paro, 1997), PH-P/PSC (Kyba and Brock, 1998), PH-P/SCM (Peterson *et al.*, 1997), PC/PSC (Kyba and Brock, 1998), PC/SCE and PSC/SCE (Gorfinkel *et al.*, 2004). In contrast, no interaction is predicted between the two subunits of the PR-DUB complex, although the N-terminal region of ASX is directly involved in its interaction with CALYPSO (Scheuermann *et al.*, 2010).

Therefore, our analysis corroborates 11 of the 17 well-established binary interactions between PCG proteins and fails to predict the remaining 6 interactions (Figures 5 and 6). In contrast, it predicts new interactions between the subunits that form the Pcl-PRC2, PRC1 and dRAF complexes, and also between subunits of PhoRC and PRC1. In addition, ASX is predicted to interact with subunits of Pcl-PRC2 complex and CALYPSO with subunits of the PhoRC, PRC1 and dRAF complexes. Further biochemical studies would be required to confirm these predicted interactions. In fact, the three subunits of dRAF (KDM2, PSC and SCE) coimmunoprecipitate (Lagarou *et al.*, 2008)

and thus they have to interact to some extent, although the directly interacting subunits and the regions involved in these interactions have not been characterized yet.

It is also noteworthy that strong interactions between subunits of the PhORC and PRC1 complexes are predicted in the coevolution analysis, which might be relevant to understand the recruitment of PRC1 at PRE. In fact, the strong evidence of coevolution detected between PHO and both PC and PH-P would indicate that these interactions have a pivotal role in the recruitment of PRC1 at PRE, in contrast to the classical view, suggesting that histone trimethylation by the E(Z) subunit of PRC2 is necessary for this recruitment. In addition, it has been proposed that SCM as a subunit of PRC1 would also be important for the recruitment of this complex at PRE (Wang *et al.*, 2010), which is consistent with the results of the coevolution analysis. The importance of PRC1 in the cooperative interactions to recruit PcG complexes at PRE has received further support in more recent studies (Kahn *et al.*, 2014). On the other hand, it should be noted that Context Mirror predicts the interactions SCM/PC, SCM/PSC and SCM/SCE, although SCM is present in substoichiometric quantities relative to the other subunits of the PRC1 complex (Wang *et al.*, 2010).

Finally, the acceleration in the nonsynonymous divergence detected in the *Bobs* branch in some genes by maximum likelihood (Figure 4) is neither necessary nor sufficient to explain the detected signals of coevolution (Figure 5). Although the higher ω in this branch (ω_{Bobs}) than in the background branches (ω_{B}) might have contributed to the predicted coevolution between subunits of the Pcl-PRC2 complex and also between SFMBT and KDM2 or PSC, other factors acting on the molecular evolution of the genes encoding these proteins in the 12 *Drosophila* species have to be likewise relevant. In fact, the partial correlations between protein evolutionary rates do not predict coevolution between SFMBT and the subunits of the Pcl-PRC2 complex despite the similar pattern of divergence in the *Bobs* branch in the genes coding these proteins and even despite the detected action of positive selection in this branch at *Sfmbt* and *E(z)*.

In summary, divergence in the *Drosophila* genus of the subunits that form the PhoRC, Pcl-PRC2, PRC1, dRAF and PR-DUB Polycomb complexes has allowed identification of putative new protein domains and the detection of episodic selection in the lineage ancestral to the



obscure group species in genes *E(z)* and *Sfinbt*, with six and seven codons, respectively, with a high posterior probability of having evolved under positive selection. In addition, the analysis of the proteomes of the species sequenced by the *Drosophila* 12 Genomes Consortium in a coevolutionary context clearly detects the footprint of coevolution not only between subunits of four Polycomb complexes, but also between subunits of different complexes. Indeed, our analysis confirms some of the well-characterized binary interactions between Polycomb proteins and predicts new interactions that deserve to be further investigated in future biochemical studies.

DATA ARCHIVING

The newly reported sequences are deposited in the EMBL/GenBank Data Libraries under accession numbers LN864764 to LN864814.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGEMENTS

We thank G Mas de Xaxars, I Salvador and NL Clark for their input in an early phase of the project, and E Puerma for her contribution in the final stages. We also thank Servei Científico-Tècnics, Universitat de Barcelona, for automated DNA sequencing facilities. This work was supported by a predoctoral fellowship from Agència de Gestió d'Ajuts Universitaris i de Recerca (AGAUR), Generalitat de Catalunya, Catalonia, Spain, to JMC-M; and grants BFU2007-63229 and BFU2012-35168 from Ministerio de Economía y Competitividad, Spain, and 2009SGR-1287 and 2014SGR10555 from Comissió Interdepartamental de Recerca i Innovació Tecnològica, Generalitat de Catalunya, Catalonia, Spain, to MA.

Alfieri C, Gambetta MC, Matos R, Glatt S, Sehr P, Fraterman S et al. (2013). Structural basis for targeting the chromatin repressor Sfinbt to Polycomb response elements. *Genes Dev* **27**: 2367–2379.

Beh LY, Colwell LJ, Francis NJ (2012). A core subunit of Polycomb repressive complex 1 is broadly conserved in function but not primary sequence. *Proc Natl Acad Sci USA* **109**: E1063–E1071.

Beisswanger S, Stephan W (2008). Evidence that strong positive selection drives neofunctionalization in the tandemly duplicated *polyhomeotic* genes in *Drosophila*. *Proc Natl Acad Sci USA* **105**: 5447–5452.

Burland TG (2000). DNASTAR's Lasergene sequence analysis software. *Methods Mol Biol* **132**: 71–91.

Clark AG, Eisen MB, Smith DR, Bergman CM, Oliver B, Markow TA et al. (2007). Evolution of genes and genomes on the *Drosophila* phylogeny. *Nature* **450**: 203–218.

Clark NL, Aquadro CF (2010). A novel method to detect proteins evolving at correlated rates: identifying new functional relationships between coevolving proteins. *Mol Biol Evol* **27**: 1152–1161.

Edgar RC (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **32**: 1792–1797.

Frankie A, Messmer S, Paro R (1995). Mapping functional domains of the Polycomb protein of *Drosophila melanogaster*. *Chromosome Res* **3**: 351–360.

Fraser HB, Hirsh AE, Wall DP, Eisen MB (2004). Coevolution of gene expression among interacting proteins. *Proc Natl Acad Sci USA* **101**: 9033–9038.

Gindhart JG, Kaufman TC (1995). Identification of *Polycomb* and *trithorax* group responsive elements in the regulatory region of the *Drosophila* homeotic gene *Sex combs reduced*. *Genetics* **139**: 797–814.

Gorfinkel N, Fanti L, Melgar T, García E, Pimpinelli S, Guerrero I et al. (2004). The *Drosophila* Polycomb group gene *Sex combs extra* encodes the ortholog of mammalian Ring1 proteins. *Mech Dev* **121**: 449–462.

Grimm C, Matos R, Ly-Hartig N, Steuerwald U, Lindner D, Rybin V et al. (2009). Molecular recognition of histone lysine methylation by the Polycomb group repressor dSfinbt. *EMBO J* **28**: 1965–1977.

Hakes L, Lovell SC, Oliver SG, Robertson DL (2007). Specificity in protein interactions and its relationship with sequence diversity and coevolution. *Proc Natl Acad Sci USA* **104**: 7999–8004.

Hauenschild A, Ringrose L, Altmutter C, Paro R, Rehmsmeier M (2008). Evolutionary plasticity of Polycomb/Trithorax response elements in *Drosophila* species. *PLoS Biol* **6**: e261.

Jones CA, Ng J, Peterson AJ, Morgan K, Simon J, Jones RS (1998). The *Drosophila* esc and *E(z)* proteins are direct partners in Polycomb group-mediated repression. *Mol Cell Biol* **18**: 2825–2834.

Jones P, Binns D, Chang H-Y, Fraser M, Li W, McAnulla C et al. (2014). InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**: 1236–1240.

Joshi P, Carrington EA, Wang L, Ketel CS, Miller EL, Jones RS et al. (2008). Dominant alleles identify SET domain residues required for histone methyltransferase of Polycomb repressive complex 2. *J Biol Chem* **283**: 27757–27766.

Juan D, Pazos F, Valencia A (2008). High-confidence prediction of global interactions based on genome-wide coevolutionary networks. *Proc Natl Acad Sci USA* **105**: 934–939.

Juan D, Pazos F, Valencia A (2013). Emerging methods in protein co-evolution. *Nat Rev Genet* **14**: 249–261.

Kahn TG, Stenberg P, Pirrotta V, Schwartz YB (2014). Combinatorial interactions are required for the efficient recruitment of Pho repressive complex (PhoRC) to Polycomb response elements. *PLoS Genet* **10**: e1004495.

Ketel CS, Andersen EF, Vargas ML, Suh J, Strome S, Simon JA (2005). Subunit contributions to histone methyltransferase activities of fly and worm Polycomb group complexes. *Mol Cell Biol* **25**: 6857–6868.

Kosakovsky Pond SLK, Murrell B, Fourment M, Frost SDW, Delpoit W, Scheffer K (2011). A random effects branch-site model for detecting episodic diversifying selection. *Mol Biol Evol* **28**: 3033–3043.

Kurzahls RL, Tie F, Stratton CA, Harte PJ (2008). *Drosophila* ESC-like can substitute for ESC and becomes required for Polycomb silencing if ESC is absent. *Dev Biol* **313**: 293–306.

Kyba M, Brock HW (1998). The *Drosophila* Polycomb group protein Psc contacts ph and Pc through specific conserved domains. *Mol Cell Biol* **18**: 2712–2720.

Lagarou A, Mohd-Sarip A, Moshkin YM, Chalkley GE, Bestarosti K, Demmers JAA et al. (2008). dKDM2 couples histone H2A ubiquitylation to histone H3 demethylation during Polycomb group silencing. *Genes Dev* **22**: 2799–2810.

Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H et al. (2007). Clustal W and Clustal X version 2.0. *Bioinformatics* **23**: 2947–2948.

Lesley Brown J, Mucci D, Whiteley M, Dirksen M-L, Kassia JA (1998). The *Drosophila* Polycomb group gene *pleiohomeotic* encodes a DNA binding protein with homology to the transcription factor YY1. *Mol Cell* **1**: 1057–1064.

Lovell SC, Robertson DL (2010). An integrated view of molecular coevolution in protein-protein interactions. *Mol Biol Evol* **27**: 2567–2575.

Löytynoja A, Goldman N (2008). Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis. *Science* **320**: 1632–1635.

Mohd-Sarip A, Venturini F, Chalkley GE, Verrijzer CP (2002). Pleiohomeotic can link Polycomb to DNA and mediate transcriptional repression. *Mol Cell Biol* **22**: 7473–7483.

Nekrasov M, Klymenko T, Fraterman S, Papp B, Oktaba K, Köcher T et al. (2007). Pcl-PRC2 is needed to generate high levels of H3-K27 trimethylation at Polycomb target genes. *EMBO J* **26**: 4078–4088.

Nowak AJ, Alfieri C, Stirnimann CU, Rybin V, Baudin F, Ly-Hartig N et al. (2011). Chromatin-modifying complex component NURF55/P55 associates with histones H3, H4 and Polycomb repressive complex 2 subunit SU(Z)12 through partially overlapping binding sites. *J Biol Chem* **286**: 23388–23396.

O'Connell S, Wang L, Robert S, Jones CA, Saint R, Jones RS (2001). Polycomblike PHD fingers mediate conserved interaction with enhancer of zeste protein. *J Biol Chem* **276**: 43065–43073.

Peterson AJ, Kyba M, Bornemann D, Morgan K, Brock HW, Simon J (1997). A domain shared by the Polycomb group proteins Scm and ph mediates heterotypic and homotypic interactions. *Mol Cell Biol* **17**: 6683–6692.

Pirrotta V, Li H-B (2012). A view of nuclear Polycomb bodies. *Curr Opin Genet Dev* **22**: 101–109.

Rychlik W (1993). Selection of primers for polymerase chain reaction. In: White BA(ed) *PCR Protocols: Current Methods and Applications*. Methods in Molecular Biology, Humana Press Inc.: Totowa, NJ, vol 15, pp 31–40.

Scheuermann JC, Gaytán de Ayala Alonso A, Oktaba K, Ly-Hartig N, McGinty RK, Fraterman S et al. (2010). Histone H2A deubiquitinase activity of the Polycomb repressive complex PR-DUB. *Nature* **465**: 243–247.

Schuettengruber B, Cavalli G (2009). Recruitment of Polycomb group complexes and their role in the dynamic regulation of cell fate choice. *Development* **136**: 3531–3542.

Schuettengruber B, Ganapathi M, Leblanc B, Portoso M, Jaschke R, Tolhuis B et al. (2009). Functional anatomy of Polycomb and Trithorax chromatin landscapes in *Drosophila* embryos. *PLoS Biol* **7**: e1000013.

Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D et al. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* **13**: 2498–2504.

Simon JA, Kingston RE (2013). Occupying chromatin: Polycomb mechanisms for getting to genomic targets, stopping transcriptional traffic, and staying put. *Mol Cell* **49**: 808–824.

Song J-J, Garlick JD, Kingston RE (2008). Structural basis of histone H4 recognition by p55. *Genes Dev* **22**: 1313–1318.

Strutt H, Paro R (1997). The Polycomb group protein complex of *Drosophila melanogaster* has different compositions at different target genes. *Mol Cell Biol* **17**: 6773–6783.

Tamura K, Stecher G, Peterson D, Filipiński A, Kumar S (2013). MEGA6: molecular evolutionary genetics analysis version 6.0. *Mol Biol Evol* **30**: 2725–2729.

Tie F, Furuyama T, Harte PJ (1998). The *Drosophila* Polycomb group proteins ESC and E(Z) bind directly to each other and co-localize at multiple chromosomal sites. *Development* **125**: 3483–3496.

Villanueva-Cañas JL, Laurie S, Albà MM (2013). Improving genome-wide scans of positive selection by using protein isoforms of similar length. *Genome Biol Evol* **5**: 457–467.

- Wang L, Brown JL, Cao R, Zhang Y, Kassisi JA, Jones RS (2004). Hierarchical recruitment of Polycomb group silencing complexes. *Mol Cell* **14**: 637–646.
- Wang L, Jähren N, Vargas ML, Andersen EF, Benes J, Zhang J *et al*. (2006). Alternative ESC and ESC-like subunits of a Polycomb group histone methyltransferase complex are differentially deployed during *Drosophila* development. *Mol Cell Biol* **26**: 2637–2647.
- Wang L, Jähren N, Miller EL, Ketel CS, Mallin DR, Simon JA (2010). Comparative analysis of chromatin binding by sex comb on midleg (SCM) and other Polycomb group repressors at a *Drosophila* Hox gene. *Mol Cell Biol* **30**: 2584–2593.
- Yang Z, Wong WSW, Nielsen R (2005). Bayes Empirical Bayes inference of amino acid sites under positive selection. *Mol Biol Evol* **22**: 1107–1118.
- Yang Z (2007). PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol* **24**: 1586–1591.
- Xue B, Dunbrack RL, Williams RW, Dunker AK, Uversky VN (2010). PONDR-FIT: a meta-predictor of intrinsically disordered amino acids. *Biochim Biophys Acta* **1804**: 996–1010.
- Zhang J, Nielsen R, Yang Z (2005). Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. *Mol Biol Evol* **22**: 2472–2479.

Supplementary Information accompanies this paper on Heredity website (<http://www.nature.com/hdy>)

Adaptive selection and coevolution at the proteins of the Polycomb repressive complexes in *Drosophila*

Calvo-Martín JM, Librado P, Aguadé M, Papaceit M, Segarra C

2016. *Heredity* 116:213–223. doi: 10.1038/hdy.2015.91

Supplementary Material

Table S1. Sequencing errors and missannotations detected in different Polycomb group genes in the species sequenced by the Drosophila 12 Genomes Consortium (Clark et al. 2007).

Gene	Species ^a	Annotation ID	Annotation and sequencing errors	Corrections
Pho	<i>D. ananassae</i>	-	Not annotated. 5' end region not sequenced.	Annotation based on the <i>D. melanogaster</i> ortholog without exons 1 and 2.
	<i>D. pseudoobscura</i>	-	Not annotated. 3' end region not sequenced.	Resequencing and annotation based on the <i>D. melanogaster</i> ortholog.
	<i>D. persimilis</i>	-	Not annotated.	Annotation based on the <i>D. melanogaster</i> ortholog.
	<i>D. mojavensis</i>	Dm9j/GI14036	Incomplete annotation. Dm9j/GI14035 belongs to Pho gene too.	Reannotation based on the <i>D. grimshawi</i> ortholog.
Elz	<i>D. virilis</i>	Dvir/GI12430	Incomplete annotation. Excludes the first two exons.	Reannotation based on the <i>D. grimshawi</i> ortholog.
	<i>D. simulans</i>	Dsim/GD12857	Three additional introns encompassing four false indels in exons 2 and 6.	Resequencing and reannotation based on the <i>D. melanogaster</i> ortholog.
	<i>D. sechellia</i>	Dsec/GM24804	Intron 2 enlarged to encompass a probably false indel in the beginning of exon 3.	Codon affected by the indel removed from the sequence to avoid phase shift. Reannotation based on the <i>D. melanogaster</i> ortholog.
	<i>D. persimilis</i>	Dper/GL15273	Premature stop codon due to the poor sequencing of the 3' end region.	Resequencing and reannotation based on the <i>D. pseudoobscura</i> ortholog.
Su(z)12	<i>D. simulans</i>	Dsim/GD12239	Exon 4 removed; exons 3 and 5 shortened and stop codon shifted due to two poorly sequenced regions.	Resequencing and reannotation based on the <i>D. melanogaster</i> ortholog.
	<i>D. persimilis</i>	Dper/GL22906	Additional intron encompassing a false indel in exon 4.	Resequencing and reannotation based on the <i>D. pseudoobscura</i> ortholog.
	<i>D. mojavensis</i>	Dmoj/GI11653	Additional intron encompassing a poorly sequenced region affecting exons 4 and 5.	Resequencing and reannotation based on the <i>D. virilis</i> ortholog.
	<i>D. mojavensis</i>	Dmoj/GI18653	Start codon shifted to CDS position 1308 to avoid a probably false indel. First exon almost entirely eliminated.	Codon affected by the indel removed from the sequence to avoid phase shift. Reannotation based on the <i>D. grimshawi</i> ortholog.
Pcl	<i>D. virilis</i>	Dvir/GJ21669	Annotation excludes the first codon.	Reannotation based on the <i>D. grimshawi</i> ortholog.
	<i>D. grimshawi</i>	Dgrn/GH21051	Additional intron in exon 5 without GT signal.	Reannotation based on the <i>D. virilis</i> ortholog.
	<i>D. simulans</i>	Dsim/GD25796	Premature CDS termination at exon 2 due to the location of exons 3 and 4 in different scaffolds, chr2R_MRandom_463 and 464. Probably false indel at the scaffold 464 affecting exon 4.	Resequencing the region between scaffolds. Codon affected by the indel removed from the sequence to avoid phase shift. Reannotation based on the <i>D. melanogaster</i> ortholog.
	<i>D. sechellia</i>	Dsec/GM20318 Dsec/GM20319	Additional intron encompassing a probably false indel in exon 1. Premature CDS termination at exon 3 due to another probably false indel with the consequent phase shift and the apparition of a stop codon that splits the gene in two annotations.	Codons affected by the indel removed from the sequence to avoid the additional intron and the split of the gene. Reannotation based on the <i>D. melanogaster</i> ortholog.
Psc	<i>D. yakuba</i>	Dyak/GE13404	Start codon shifted to CDS position 94.	Reannotation based on the <i>D. melanogaster</i> ortholog.
	<i>D. erecta</i>	Dere/GG22533	Start codon shifted to CDS position 94.	Reannotation based on the <i>D. melanogaster</i> ortholog.
	<i>D. pseudoobscura</i>	Dpse/GA17752	Start codon shifted to CDS position 61 and additional intron in exon 4.	Reannotation based on the <i>D. melanogaster</i> ortholog.
	<i>D. persimilis</i>	Dper/GL10594	Start codon shifted to CDS position 55 and additional intron in exon 4.	Reannotation based on the <i>D. melanogaster</i> ortholog.
	<i>D. willistoni</i>	Dwil/GK20868	Additional intron in exon 4.	Reannotation based on the <i>D. melanogaster</i> ortholog.
	<i>D. mojavensis</i>	Dmoj/GI19568	Start codon shifted to CDS position 286 and additional intron in exon 4.	Reannotation based on the <i>D. melanogaster</i> ortholog.
	<i>D. virilis</i>	Dvir/GJ21143	Start codon shifted to CDS position 37.	Reannotation based on the <i>D. melanogaster</i> ortholog.
	<i>D. grimshawi</i>	Dgrn/GH21599	Start codon shifted to CDS position 757.	Reannotation based on the <i>D. melanogaster</i> ortholog.
	<i>D. grimshawi</i>	Dgrn/GH17005	Additional intron in exon 1.	Reannotation based on the <i>D. melanogaster</i> ortholog.
	<i>D. sechellia</i>	Dsec/GM22147	Additional intron encompassing a poorly sequenced region in exon 2.	129 bp region removed from the sequence to avoid all codons affected.
	<i>D. mojavensis</i>	Dmoj/GI13310	Additional intron in exon 2 without GT signal.	Reannotation based on the <i>D. melanogaster</i> ortholog.
	<i>D. virilis</i>	Dvir/GI12080	Additional intron in exon 2 without GT signal.	Reannotation based on the <i>D. melanogaster</i> ortholog.

Gene	Species ^a	Annotation ID	Annotation and sequencing errors	Corrections
Ph-p	<i>D. simulans</i>	Dsim(GD16391)	Only Dsim(GD16391 and Dsim(GD24614 corresponds to the Ph-p gene, Dsim(GD13792 corresponds to its paralog Ph-d. The gene is splitted in two different annotations due to two unsequenced regions.	Resequencing and reannotation based on the <i>D. melanogaster</i> ortholog.
		Dsim(GD24614)		
	<i>D. sechellia</i>	Dsec(GM18916)	Only Dsec(GM18916 corresponds to the Ph-p gene, Dsec(GM18919 corresponds to its paralog Ph-d. A large fraction of the gene is unsequenced.	Resequencing and reannotation based on the <i>D. melanogaster</i> ortholog.
		Dsec(GM18919)		
<i>D. yakuba</i>	<i>D. yakuba</i>	Dyak(GE16976)	Only Dyak(GE16976 corresponds to the Ph-p gene, Dyak(GE16977 corresponds to its paralog Ph-d. Small region in the exon 2 unsequenced.	Resequencing and reannotation based on the <i>D. melanogaster</i> ortholog.
		Dyak(GE16977)		
<i>D. erecta</i>	<i>D. erecta</i>	Dere(GG12650)	Ph-p corresponds to the 5' half of Dere(GG12650 and the 3' half of Dere(GG12564, due to a bad assembling with the sequences of its paralog Ph-d.	Reannotation based on the sequence similarities with the <i>D. yakuba</i> genes.
		Dere(GG12564)		
<i>D. ananasae</i>	<i>D. ananasae</i>	Dana(GF21461)	Only Dana(GF21461 corresponds to the Ph-p gene, Dana(GF20806 corresponds to its paralog Ph-d. Additional intron in exon 4 without GT signal.	Reannotation based on the <i>D. melanogaster</i> ortholog.
		Dana(GF20806)		
<i>D. pseudoobscura</i>	<i>D. pseudoobscura</i>	Dpse(GA22279)	Only Dpse(GA22279 corresponds to the Ph-p gene, Dpse(GA22280 corresponds to its paralog Ph-d.	No corrections.
		Dpse(GA22280)		
<i>D. persimilis</i>	<i>D. persimilis</i>	Dper(GL15368)	Both annotations correspond to fragments of Ph-d. Ph-p is not annotated and almost entirely unsequenced.	Resequencing and reannotation based on the <i>D. pseudoobscura</i> ortholog.
		Dper(GL23489)		
<i>D. willistoni</i>	<i>D. willistoni</i>	Dwil(GK24927)	Exon 1 of Ph-p not annotated and replaced by an exon similar to the exon 1 in Ph-d, although seems that in <i>D. willistoni</i> the paralog Ph-d don't exist.	Reannotation based on the <i>D. melanogaster</i> ortholog.
<i>D. mojavensis</i>	<i>D. mojavensis</i>	Dmoj(GI11060)	5 additional introns, only one with GT signal.	Reannotation based on the <i>D. melanogaster</i> ortholog.
		Dvir(GI16511)	Additional intron without GT signal in the exon 2 to avoid a poorly sequenced region. Premature stop codon in the exon 5, leaving out of the gene a hyperconserved 3' end region, probably due to another poorly sequenced region.	Poorly sequenced regions excluded from the analysis. Hyperconserved 3' end region reincluded in the CDS sequence. Reannotation based on the <i>D. melanogaster</i> ortholog.
Scm	<i>D. grimshawi</i>	Dgr(GH24401)	3 additional introns, only one with GT signal. Start of the exon 5 delayed 113 bp.	Reannotation based on the <i>D. melanogaster</i> ortholog.
	<i>D. simulans</i>	Dsim(GD20776)	5' extreme of the exon 1 not sequenced. 3' extreme of the exon 1, whole exon 2 and beginning of the exon 3 not annotated and located apart in scaffold chru_M_6542.	Resequencing and reannotation based on the <i>D. melanogaster</i> ortholog.
		Dsec(GM26235)	Exon 1 and part of exon 2 annotated apart as Dsec(GM26234 to avoid a probably false indel in the beginning of exon 2 and the resulting phase shift.	Codon affected by the indel removed from the sequence to avoid the phase shift and the split of the gene. Reannotation based on the <i>D. melanogaster</i> ortholog.
Kdm2	<i>D. sechellia</i>	-	Not annotated. Erroneously included in the annotation Dper(GL12262 corresponding to the adjacent gene.	Reannotation based on the <i>D. pseudoobscura</i> ortholog.
	<i>D. simulans</i>	Dsim(GD18594)	Part of the exon 4 annotated apart as Dsim(GD17761 due to a false indel after the AG signal of the intron 3 and the resulting phase shift.	Resequencing and reannotation based on the <i>D. melanogaster</i> ortholog.
		Dyak(GE25927)	Part of the exon 4 annotated apart as Dyak(GE25928 without apparent cause.	Reannotation based on the <i>D. melanogaster</i> ortholog.
Calypro	<i>D. yakuba</i>	Dana(GF16580)	3' end region with phase shift and premature stop codon due to a probably false indel.	Codon affected removed from the sequence to avoid phase shift and reannotation based on the <i>D. melanogaster</i> ortholog.
	<i>D. ananasae</i>			
Asx	<i>D. willistoni</i>	Dwil(GK22163)	Start codon shifted to CDS position 136.	Reannotation based on the <i>D. melanogaster</i> ortholog.
	<i>D. sechellia</i>	Dsec(GM21552)	Premature stop codon causing loss of the two last exons due to the poor sequencing of the 3' end region.	Resequencing and reannotation based on the <i>D. melanogaster</i> ortholog.
		Dper(GL11443)	Premature stop codon causing loss of the last exon.	Reannotation based on the <i>D. melanogaster</i> ortholog.
<i>D. persimilis</i>	<i>D. persimilis</i>	Dmoj(GI20581)	Two additional introns without GT-AG signal.	Reannotation based on the <i>D. melanogaster</i> ortholog.
	<i>D. mojavensis</i>			
<i>D. grimshawi</i>	<i>D. grimshawi</i>	Dgr(GH22785)	Premature stop codon causing loss of the last exon.	Reannotation based on the <i>D. melanogaster</i> ortholog.

^a Genomic DNA (gDNA) of *D. simulans*, *D. yakuba*, *D. pseudoobscura* and *D. persimilis* were available in our laboratory. Strains of *D. mojavensis* and *D. sechellia* were kindly provided by A. Ruiz (UAB, Barcelona) and M. Louis (CRG, Barcelona), respectively.

Artículo 2

Evidence of neofunctionalization after the duplication of the highly conserved Polycomb group gene *Caf1-55* in the obscura group of *Drosophila*

Calvo-Martín JM, Papaceit M, Segarra C

2017. *Scientific Reports* 7:40536. doi: 10.1038/srep40536

La proteína CAF1-55 en el género *Drosophila* constituye una de las subunidades del complejo represor Polycomb 2 (PRC2) y de otros complejos multiméricos. Esta proteína, multifuncional y evolutivamente conservada, participa en el ensamblaje y remodelación de los nucleosomas y en la regulación epigenética de un gran conjunto de genes diana. En este trabajo describimos y analizamos la duplicación del gen *Caf1-55* en las especies del grupo *obscura* de *Drosophila*. Ambos parálogos exhiben una fuerte asimetría en sus tasas evolutivas, lo cual sugiere que han evolucionado de acuerdo a un proceso de neofuncionalización. Durante dicho proceso, la copia ancestral se ha mantenido conservada bajo una fuerte selección purificadora, reteniendo su función original, mientras que la copia derivada (*Caf1-55dup*), originada en un evento de duplicación mediado por ADN hace ~18 millones de años, ha evolucionado por episodios de selección adaptativa. Diversos métodos de máxima verosimilitud confirman la acción de la selección positiva sobre *Caf1-55dup* tras la duplicación, a la vez que descartan una relajación de la selección. Este proceso adaptativo también ha tenido lugar a una escala más reciente durante la divergencia de *D. subobscura* y *D. guanche*. Se discute la posible relación de esta duplicación con la aceleración de las tasas evolutivas en el linaje ancestral del grupo *obscura* de otras tres proteínas integrantes del complejo PRC2. Por último, se compara el momento de origen y las consecuencias funcionales de la duplicación de *Caf1-55* con las duplicaciones de otros genes Polycomb.

SCIENTIFIC REPORTS

OPEN

Evidence of neofunctionalization after the duplication of the highly conserved Polycomb group gene *Caf1-55* in the obscura group of *Drosophila*

Received: 02 August 2016
Accepted: 07 December 2016
Published: 17 January 2017

Juan M. Calvo-Martin, Montserrat Papaceit & Carmen Segarra

Drosophila CAF1-55 protein is a subunit of the Polycomb repressive complex PRC2 and other protein complexes. It is a multifunctional and evolutionarily conserved protein that participates in nucleosome assembly and remodelling, as well as in the epigenetic regulation of a large set of target genes. Here, we describe and analyze the duplication of *Caf1-55* in the obscura group of *Drosophila*. Paralogues exhibited a strong asymmetry in evolutionary rates, which suggests that they have evolved according to a neofunctionalization process. During this process, the ancestral copy has been kept under steady purifying selection to retain the ancestral function and the derived copy (*Caf1-55dup*) that originated via a DNA-mediated duplication event ~18 Mya, has been under clear episodic selection. Different maximum likelihood approaches confirmed the action of positive selection, in contrast to relaxed selection, on *Caf1-55dup* after the duplication. This adaptive process has also taken place more recently during the divergence of *D. subobscura* and *D. guanche*. The possible association of this duplication with a previously detected acceleration in the evolutionary rate of three CAF1-55 partners in PRC2 complexes is discussed. Finally, the timing and functional consequences of the *Caf1-55* duplication is compared to other duplications of Polycomb genes.

Drosophila CAF1-55 protein (also known as CAF1, p55 and NURF55) is part of the Polycomb repressive complex PRC2^{1,2}. This protein complex is responsible for the trimethylation of the histone H3 lysine 27 (H3K27me3) that is one of the posttranslational histone modifications introduced by the Polycomb repressive complexes at specific target sites to modulate the chromatin state. However, CAF1-55 is a multifunctional protein and is a subunit of other protein complexes such as CAF1, NURF, NuRD and REAM/MMB. The heterotrimeric complex CAF1 (chromatin assembly factor 1) is involved in the assembly of nucleosomes after DNA replication³. NURF (nucleosome remodelling factor) is an ATP-dependent chromatin remodelling complex⁴ and NuRD is a nucleosome remodelling and deacetylase complex⁵. REAM (Rb, E2F and Myb complex) and MMB (Myb-MuvB complex) are two similar protein complexes that have been independently purified and participate in the activation and repression of developmental genes and origins of DNA replication^{6,7}.

A common function of CAF1-55 in all these complexes is to serve as a scaffold to facilitate the interaction between histones and other proteins. Indeed, the interaction of CAF1-55 with the first helix of histone H4 and other proteins has been resolved by crystal structure analysis⁸. Structural analysis indicates that CAF1-55 is a noncatalytic protein and a member of the WD40 family with a seven-bladed β -propeller structure. WD40 proteins participate in protein-protein interactions, and they are overrepresented among proteins involved in interactome networks⁹. CAF1-55 interacts by means of the WD40 repeats with 35 different *D. melanogaster* proteins, as reported in the BioGRID database¹⁰.

The pivotal role of CAF1-55 in chromatin metabolism, as well as its ability to interact with a wide range of proteins, indicates that CAF1-55 is a hub protein with multiple pleiotropic effects, which makes it an essential

Departament de Genètica, Microbiologia i Estadística, Facultat de Biologia, i Institut de Recerca de la Biodiversitat (IRBio), Universitat de Barcelona, Barcelona, Spain. Correspondence and requests for materials should be addressed to C.S. (email: csegarra@ub.edu)

protein. Indeed, *Caf1-55* null alleles cause lethality before pupariation and mutant larvae die mostly at the second instar stage¹¹. Reduced levels of *Caf1-55* expression result in homeotic transformations likely due to misregulation of the Hox genes by PRC2¹². Proteins with a large number of interactors, especially if they are located at the center of a network, are subject to strong constraints on variation and are evolutionarily conserved¹³, which is consistent with the presence of CAF1-55 in a wide range of species, from fungi to mammals and plants, and with its high level of amino acid conservation. In fact, the sequence identity between the *Drosophila melanogaster* CAF1-55 and the human homologs RbAp48 and RbAp46 is 87% and 86%, respectively³.

In a study of the molecular evolution of *Caf1-55* and other Polycomb group (PcG) genes in *Drosophila*¹⁴, the orthologs of *Caf1-55* were identified in a set of 15 species, and it was confirmed that CAF1-55 is a highly conserved protein with minimal interspecific amino acid divergence. In the same study, a gene with a high similarity with *Caf1-55* was unexpectedly detected in *D. pseudoobscura* and *D. persimilis*. This gene (henceforth, *Caf1-55dup*) was absent in the other *Drosophila* species sequenced by the *Drosophila* 12 Genomes Consortium¹⁵. This result suggested the duplication of *Caf1-55* prior to the divergence of the obscura group species. Herein, we analyze the molecular evolution of *Caf1-55* and *Caf1-55dup* paralogs both at the interspecific level in the *Drosophila* genus and at the intraspecific level in *D. subobscura* in order to infer the evolutionary history of this duplication. Different alternative scenarios can drive the evolution of duplicated genes: conservation (both copies retain the ancestral function in a redundant manner), nonfunctionalization (one copy retains the ancestral function and the other becomes silenced and degenerates), neofunctionalization (one copy retains the ancestral function and the other acquires a new function), subfunctionalization (both copies retain a different subset of the ancestral functions) and specialization (both copies acquire a novel function different from the ancestral one). Among these scenarios, only neofunctionalization and specialization permit the origin of new genes and functions¹⁶. Thus, the first aim of this study was to disentangle under which of these scenarios did *Caf1-55* and *Caf1-55dup* evolve. In fact, purifying selection acting on the ancestral gene may be relaxed after the duplication due to gene redundancy. This relaxation might have reduced the extent of conservation of CAF1-55. Moreover, this duplication might have increased the dosage of CAF1-55, which might have been a challenge for a protein involved in different complexes if it led to an imbalance in the concentration of the different subunits that form these complexes¹⁷.

The results obtained indicate that the duplication of *Caf1-55* occurred about 18 Mya in the lineage ancestral to the obscura group species and that both paralogs likely evolved under a neofunctionalization process, in which strong purifying selection was maintained on the ancestral *Caf1-55* gene and positive selection acted on the new *Caf1-55dup* gene. The action of positive selection on *Caf1-55dup* was not only detected immediately after its origin, but also more recently, specifically since the divergence of *D. subobscura* and *D. guanche*. In addition, the results show that *Caf1-55* is a dynamic gene, as it underwent at least an additional duplication event in the *D. persimilis* lineage (about 0.35 Mya).

Results

Identification of the *Caf1-55* orthologs and paralogs. *Caf1-55* orthologs in the species sequenced by the *Drosophila* 12 Genomes Consortium¹⁵ are described in FlyBase (www.flybase.org). However, BLASTN searches using as query the coding region of the *Caf1-55* gene of *D. melanogaster* (CG4236) were performed to corroborate the available data. The sequences with the highest similarity to the query and with E values close to 0 were GA18051 in *D. pseudoobscura* and GL12530 in *D. persimilis*. Synteny with flanking genes is conserved when comparing CG4236, GA18051 and GL12530, which confirmed that the three genes are orthologs. Unexpectedly, these BLASTN searches retrieved other genes with a rather high similarity to *Caf1-55* in *D. pseudoobscura* (GA26389; E value = 6.48×10^{-12}) and *D. persimilis* (GL21757; E value = 2.47×10^{-8}). These two genes are located in a conserved syntenic region that is different from that of *Caf1-55*. According to FlyBase, GA26389 or GL21757 have not orthologs in the other species sequenced by the *Drosophila* 12 Genomes Consortium. Additional BLASTN searches using as query these genes and analysis of the syntenic region where they are located corroborated that GA26389 or GL21757 are absent in these species. This result suggested a duplication of *Caf1-55* in the obscura group species, which was further confirmed by the presence of the *Caf1-55* duplicate, *Caf1-55dup*, in the available genome of *D. miranda*¹⁸ and by the successful PCR amplification and sequencing of *Caf1-55dup* in three additional species of the obscura group (*D. subobscura*, *D. madeirensis* and *D. guanche*). It is remarkable that *Caf1-55* has an additional duplicate (GL12106) in the annotated *D. persimilis* genome, which is misidentified as a *Caf1-55* ortholog in FlyBase. The presence of this paralog was confirmed by its PCR amplification and sequencing in a *D. persimilis* line available in our laboratory.

In *D. pseudoobscura*, *D. persimilis* and *D. miranda* *Caf1-55* and *Caf1-55dup* are located on chromosome 2 (Muller's element E), about 8 Mb apart. *In situ* hybridization confirmed the location of both genes in the same chromosomal element (chromosome O) of *D. subobscura* (see Supplementary Fig. S1), in a region with a recombination rate of about 5 cM/Mb¹⁹. *Caf1-55dup* has three exons and two introns in the six obscura species and thus has kept the same organization of the ancestral gene, indicating a DNA-mediated duplication event. In *D. persimilis*, the annotated *Caf1-55dup* gene (GL21757) lacks the first exon. However, sequence homology indicates that this exon is present in the genomic sequence upstream from the gene, which clearly suggests a misannotation. In *D. pseudoobscura* and *D. persimilis*, *Caf1-55dup* is a nested gene inserted in the fourth intron of GA27362 and GL22062, respectively, these being orthologs of the *D. melanogaster* gene *dpr11* (CG33202). Moreover, the multiple alignment of the sequenced gene regions of *Caf1-55dup* in *D. subobscura*, *D. madeirensis* and *D. guanche* with the sequence of the fourth intron of *dpr11* in *D. melanogaster* shows regions with a clear homology flanking *Caf1-55dup*. Therefore, the genomic location of *Caf1-55dup* is maintained in the obscura group species, which indicates that this is its ancestral position at least prior to the divergence of these species.

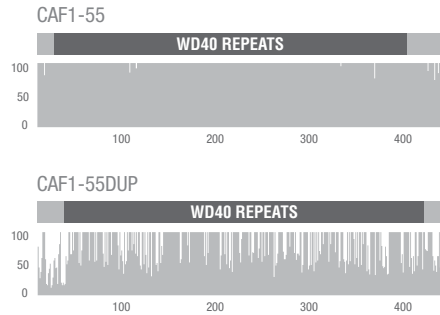


Figure 1. Protein conservation plots across CAF1-55 and CAF1-55DUP according to interspecific divergence. x-axis, amino acid sites along the multiple alignment of each protein. y-axis, inferred amino acid conservation score at each site. The grey bar above each plot shows the described WD40 domain in a black box.

The expression analysis (not including *D. miranda*) confirmed that *Caf1-55dup* is transcribed in the adult stage of the obscura group species (see Supplementary Fig. S2). Moreover, the *Caf1-55dup* cDNA sequences that were obtained lacked introns, supporting the notion that the recovered cDNA was from processed and likely functional mRNA.

Divergence of *Caf1-55* and its paralogs. The *Caf1-55* and *Caf1-55dup* sequences retrieved from FlyBase were aligned with those reported in Calvo-Martin *et al.*¹⁴ and in the present study. This alignment was based on a multiple alignment of the encoded proteins (see Supplementary Fig. S3), which included 23 sequences (16 CAF1-55, 6 CAF1-55DUP and the protein encoded by GL12106 that is the additional annotated *D. persimilis* duplicate of *Caf1-55*). Amino acid conservation is much higher along CAF1-55 than along CAF1-55DUP (Fig. 1) suggesting that purifying selection is stronger in *Caf1-55* than in its paralog. *Caf1-55dup* lacks the start codon in *D. miranda*, which was further confirmed by sequencing this gene in a line of this species available in our laboratory. Therefore, in *D. miranda* *Caf1-55dup* might be undergoing a process of pseudogenization or, alternatively, a different ATG may be in use as the start codon to translate the gene. In fact, an ATG in the first intron in frame with the second exon could render a CAF1-55DUP protein that only differs in the first five amino acids relative to that of *D. pseudoobscura* and *D. persimilis*. However, given this ambiguity and the possibility that *Caf1-55dup* is a pseudogene in *D. miranda*, this species was excluded from all subsequent analyses. On the other hand, the alignment of the *Caf1-55* and *Caf1-55dup* 5' flanking regions in the obscura group species revealed that sequence homology extends only to around the transcription start site and thus that the duplicated region includes the *Caf1-55* 5' UTR but not the upstream sequences. The absence of the promoter and other regulatory sequences in the duplicated region suggests that current expression of *Caf1-55dup* is directed by newly arisen *cis*-acting regulatory elements.

The maximum likelihood tree inferred from the multiple nucleotide alignment shows a strong increase in the substitution rate after the duplication event that gave rise to *Caf1-55dup* (Fig. 2). As a first step towards analyzing this result, three evolutionary branch models implemented in PAML²⁰ were compared to detect the presence of lineages with a significant increase in the ω estimates ($\omega = d_N/d_S$) in the phylogeny. The 3R model, which assumes three different ω estimates (one for the branches of the ancestral *Caf1-55* gene, one for *Caf1-55dup* and one for the unique *D. persimilis* duplicate GL12106), fitted the data better than the null model M0, which assumes a single ω for all branches ($p < 0.0001$). Likewise, the null model M0 was rejected when compared with the FR model, which has a different ω for each branch ($p < 0.0001$). In contrast, the FR model did not explain the data better than 3R when this was used as the null model ($p = 0.1229$). These results clearly indicate that the heterogeneity in ω values among branches is mainly due to differences in ω in the branches of the ancestral *Caf1-55*, *Caf1-55dup* and GL12106. In fact, $\omega_{Caf1-55} = 0.0012$, $\omega_{Caf1-55dup} = 0.3787$ and $\omega_{GL12106} = 1.3842$. Different likelihood methods have been developed to infer whether increases in ω estimates in particular lineages can be explained by the action of positive selection in these lineages. The aBSREL random effects branch site method²¹ provided evidence of episodic selection in three *Caf1-55dup* lineages (Fig. 2): the *Caf1-55dup* lineage prior to the divergence of the obscura group species ($p = 0.0037$), the lineage prior to the divergence of *D. pseudoobscura* and *D. persimilis* ($p = 0.0058$) and the lineage prior to the divergence of *D. subobscura* and *D. madeirensis* ($p = 0.0193$). However, p -values remained significant only for the *Caf1-55dup* lineage that leads to the obscura species after correcting for multiple testing ($p = 0.0412$). These results were also confirmed by the branch site models implemented in PAML, showing positive selection in the lineage prior to the divergence of the obscura group species ($p = 0.0002$).

Although the aBSREL and PAML branch site models were developed to detect positive selection, any increase in ω estimates may also be due to relaxed purifying selection. The RELAX method²², which relies on the aBSREL random effects branch site model, was developed to distinguish between positive and relaxed selection. This method was applied to our data set, considering the *Caf1-55* branches as reference branches and the *Caf1-55dup* branch prior to the divergence of the obscura group species as the test branch. The action of positive diversifying selection was confirmed in this last lineage ($p < 0.0001$) with a selection intensity parameter (k) equal to 50.

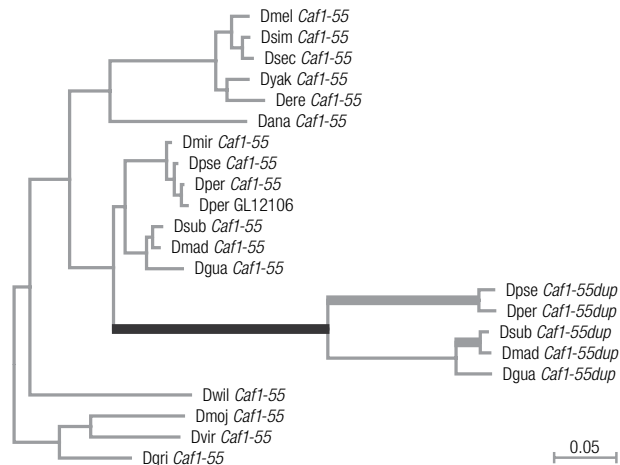


Figure 2. Phylogenetic tree based on the divergence of *Caf1-55* and *Caf1-55dup*. The three thick branches correspond to those branches with evidence of positive selection (p -value < 0.05) according to the aBSREL random effects branch site method²¹. The thick black branch remains significant after correction for multiple testing. The scale in the lower right corner indicates nucleotide substitutions per site. *Dmel* = *D. melanogaster*, *Dsim* = *D. simulans*, *Dsec* = *D. sechellia*, *Dyak* = *D. yakuba*, *Dere* = *D. erecta*, *Dana* = *D. ananassae*, *Dmir* = *D. miranda*, *Dpse* = *D. pseudoobscura*, *Dper* = *D. persimilis*, *Dsub* = *D. subobscura*, *Dmad* = *D. madeirensis*, *Dgua* = *D. guanche*, *Dwil* = *D. willistoni*, *Dmoj* = *D. mojavensis*, *Dvir* = *D. virilis* and *Dgri* = *D. grimshawi*.

The Bayes Empirical Bayes (BEB) approach²³ implemented in PAML identified 22 codons with a high posterior probability of having evolved under positive selection (see Supplementary Fig. S3) in *Caf1-55dup* after its origin by duplication. Most of these codons also show a rather high evidence ratio of having been the target of positive selection according to BUSTED²⁴.

It is noteworthy that despite the strong acceleration in the substitution rate detected in the *Caf1-55dup* lineage after the duplication event, no effect in the *Caf1-55* lineage can be inferred by either the PAML or aBSREL likelihood methods. Therefore, there is no evidence of any relaxation of the purifying selection acting on the ancestral *Caf1-55* gene as a consequence of its duplication. In fact, there are no fixed amino acid differences in *CAF1-55* between species of the obscura group and other *Drosophila* species (see Supplementary Fig. S3).

The results of the WDSP^{25,26} analysis indicated that *CAF1-55DUP* contains the seven WD40 repeats present in *CAF1-55*. The average scores estimated by WDSP for the WD40 repeats of *CAF1-55* are 86.70 and 86.57 in *D. melanogaster* and *D. pseudoobscura*, respectively. These scores for *CAF1-55DUP* are also high in the five species of the obscura group, ranging from 83.36 in *D. persimilis* to 87.48 in *D. subobscura*. The first repeat has the lowest score in both paralogs and is slightly displaced in the *CAF1-55DUP* protein of *D. subobscura*, *D. madeirensis* and *D. guanche* (see Supplementary Fig. S3). On the other hand, *CAF1-55DUP* contains 21 of the 22 residues that in *CAF1-55* are involved in the formation of the hydrogen bonds that shape the four β -strands of each blade present in the seven-bladed β -propeller structure. However, the residues potentially implicated in protein–protein interactions (hotspots) are more divergent between the paralogs (Fig. 3). Residues that have evolved under positive selection according to the BEB and BUSTED methods are mainly located in the A–B or C–D loops and D β -strands (see Supplementary Fig. S3).

The duplication of *Caf1-55* in the obscura lineage would have occurred 18 Mya (95% highest posterior density (HPD) interval: 13–23 Mya) according to the BEAST 2 software²⁷. This estimate is relative to the calibration points used in the analysis²⁸, which assumed that the *Drosophila* and *Sophophora* subgenera split 32 Mya and that the *melanogaster* and *obscura* groups split 24 Mya. In addition, the origin of the additional *Caf1-55* duplication in the *D. persimilis* lineage would have taken place 0.35 Mya (95% HPD interval: 0.05–0.76 Mya). This estimate is consistent with the estimated 0.50 Myr of divergence between *D. pseudoobscura* and *D. persimilis*²⁹.

***Caf1-55* and *Caf1-55dup* nucleotide polymorphism in *D. subobscura*.** Nucleotide polymorphism in *D. subobscura* was analyzed in *Caf1-55* and *Caf1-55dup*. The multiple alignment of the 14 *Caf1-55* and 16 *Caf1-55dup* sequences identified 85 and 95 nucleotide polymorphic sites, respectively (see Supplementary Fig. S4). A *Caf1-55dup* polymorphic site in line OF14 is a nonsense mutation that causes the loss of the last three amino acids in the encoded protein. Nucleotide diversity ($\pi = 0.0082$ and $\pi = 0.0081$ in *Caf1-55* and *Caf1-55dup*, respectively) is similar in both genes (Table 1). The pattern of variation indicates an excess of singletons (although not

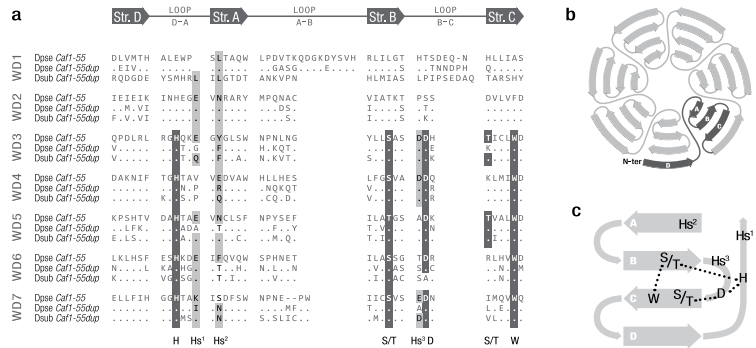


Figure 3. Comparison of the seven WD40 repeats between Cafi1-55 and Cafi1-55dup. (a) Alignment of the seven WD40 repeats, which are composed by four β -strands (Str.) interspersed with loops, as represented graphically above the alignment. Loop C–D is not included due to its poor conservation and absence of relevant residues. The ancestral Cafi1-55 protein is represented by the sequence of *D. pseudoobscura* and the Cafi1-55dup protein by the sequences of *D. pseudoobscura* and *D. subobscura*. Highlighted in black are conserved residues involved in the formation of hydrogen bonds (His, Ser/Thr, Asp, Trp) and highlighted in grey are hotspot residues (Hs¹⁻³) implicated in protein–protein interactions (i.e., any of the binding-type amino acids: Arg, His, Lys, Asp, Glu, Trp, Tyr, Phe, Leu, Ile, Met, Asn or Gln). (b) Diagram of the general β -propeller structure. The first WD40 repeat is highlighted in black, showing that it is composed by the D strand of one blade and the A, B and C strands of the next blade. (c) Location of the key residues of each repeat over a β -propeller blade. Hydrogen bonds are represented by dotted lines.

	Cafi1-55	Cafi1-55dup
number of sequences	14	16
number of sites	2658	2504
number of polymorphic sites (S)	85	95
number of singletons	46	58
nucleotide diversity (π)	0.0082	0.0081
Tajima's D	−0.8618	−1.3430
synonymous diversity (π_s)	0.0208	0.0222
nonsynonymous diversity (π_n)	0	0.0022
synonymous divergence (K_s)	0.1417	0.1009
nonsynonymous divergence (K_n)	0.0010	0.0327
K_n/K_s	0.0070	0.3238

Table 1. Estimates of nucleotide polymorphism in *D. subobscura* and of divergence between *D. subobscura* and *D. guanche* corrected for multiple hits.

significant) as reflected by the negative sign of Tajima's D³⁰ statistic. This would be expected as *D. subobscura* shows a genomewide excess of low frequency variants likely due to a population expansion soon after the penultimate glacial period³¹. Levels of synonymous variation in the coding region are also similar in *Cafi1-55* and its paralog *Cafi1-55dup* ($\pi_s = 0.0208$ in *Cafi1-55* and $\pi_s = 0.0222$ in *Cafi1-55dup*). In contrast, both genes differ substantially in levels of nonsynonymous variation. In fact, no nonsynonymous polymorphism was detected in *Cafi1-55*, whereas *Cafi1-55dup* presents 12 polymorphisms ($\pi_n = 0.0022$) that affect the encoded protein (see Supplementary Fig. S4). This difference in nonsynonymous variation is also evident in the estimates of nonsynonymous divergence between *D. subobscura* and *D. guanche* using either K_n or the K_n/K_s ratio (Table 1). These results clearly indicate much stronger functional constraints and thus purifying selection acting against nonsynonymous substitutions in *Cafi1-55* than in *Cafi1-55dup*.

The HKA test³² did not detect a decoupling between silent polymorphism and divergence when comparing *Cafi1-55* and *Cafi1-55dup* genes ($\chi^2 = 0.4167$, 1 df, $p = 0.518$) using *D. guanche* as the outgroup. On the other hand, the MK test³³ was used independently for each gene (*Cafi1-55* or *Cafi1-55dup*) to detect a putative decoupling in the polymorphism to divergence ratio for synonymous and nonsynonymous mutations. The MK test rendered a significant result only in *Cafi1-55dup*. In fact, the number of synonymous and nonsynonymous polymorphisms (24 and 12, respectively) and the number of synonymous and nonsynonymous fixed differences (21 and 30, respectively) differed significantly according to a χ^2 test of independence ($\chi^2 = 5.49$, 1 df, $p = 0.0191$). This result

and a neutrality index lower than 1 ($NI = 0.350$) indicate a significant excess of nonsynonymous substitutions fixed in *Caf1-55dup* during the divergence of *D. subobscura* and *D. guanche*. The fraction of adaptive amino acid substitutions estimated according to the α parameter³⁴ is 0.650. The α parameter might be overestimated and even evidence of adaptive selection inferred by the MK test might be artifactual when nonsynonymous mutations are under weak selection and there are strong differences in effective size between the ancestral and current populations^{35,36}. This is because slightly deleterious nonsynonymous mutations might have been fixed in a small ancestral population but they no longer segregate in a current large population. The demographic history of *D. subobscura* indicates that the species is under an expansion process³¹. However, no evidence of an important reduction in effective population size in the past was inferred. Therefore, it seems unlikely that the *D. subobscura* changes in effective size might have biased the results of the MK test.

Discussion

Gene duplication is an important evolutionary mechanism to generate new genes and functions. Different models have been proposed to explain the origin, maintenance and evolution of gene duplicates³⁷. In *Drosophila*, the most prevalent mechanism by which duplicated genes are retained is neofunctionalization¹⁶. Under neofunctionalization one gene is under purifying selection and retains the ancestral function, whereas the other gene acquires a new function by positive selection. The differential action of selection on both paralogs causes a strong asymmetry in nonsynonymous substitution rates³⁸. Generally, the ancestral gene is the constrained copy that evolves under purifying selection and the new duplicate is the unconstrained copy that acquires a new function under positive selection^{39,40}. The disruption in the new copy of the regulatory sequences, mainly of the 5' flanking region, during the duplication process itself can explain how selection discriminates between the two paralogs⁴¹. Despite this general scenario, relaxation of purifying selection during divergence between paralogs cannot be ruled out either for the ancestral copy due to gene redundancy or the new gene.

The evolution of the *Caf1-55* and *Caf1-55dup* paralogs analyzed herein seems to have followed a neofunctionalization process. This process has been led by strong purifying selection acting on the ancestral copy (*Caf1-55*) and by positive selection on the new copy (*Caf1-55dup*). In fact, there is no evidence of relaxed selection in *Caf1-55* after the duplication, which means that selection could have discriminated between the redundant gene copies immediately after their origin. The *Caf1-55* duplication occurred via a DNA-mediated event and only affected *Caf1-55*, as those genes flanking *Caf1-55* are absent around *Caf1-55dup*. Moreover, the duplicated segment likely did not include the regulatory regions upstream from *Caf1-55* since sequence similarity when comparing *Caf1-55* and *Caf1-55dup* in the obscure species is restricted to the 5' UTR. It is thus feasible that *Caf1-55dup* was not actually a redundant version of *Caf1-55* after its origin because it was poorly expressed or was not expressed at all. This argument, however, prompts the question why *Caf1-55dup* was maintained and fixed. It is possible that the expression level of *Caf1-55dup* increased only after the gene had accumulated enough mutations to ensure a difference in function between *CAF1-55* and *CAF1-55DUP*. In contrast to the strong purifying selection acting on *Caf1-55*, there is clear evidence of the action of adaptive positive selection on *Caf1-55dup*. The adaptive process can be detected both in the lineage ancestral to the obscure group species (i.e., after the duplication), and more recently during the divergence of *D. subobscura* and *D. guanche*, as revealed by the MK test in the intra- and inter-specific analysis. In addition, the intraspecific analysis indicated that purifying selection against nonsynonymous polymorphism is much stronger in *Caf1-55* than in *Caf1-55dup* (Table 1).

CAF1-55 is a subunit of the PRC2 Polycomb complex. This complex contains three additional proteins: *E(Z)*, *ESC* and *SU(Z)12*. In a previous study¹⁴, it was shown that the genes coding these proteins suffered a significant increase in the nonsynonymous substitution rate in the lineage ancestral to the obscure group species. Therefore, in this lineage not only took place the duplication of *Caf1-55*, but also an acceleration in the fixation rate of nonsynonymous changes in the genes encoding proteins that interact with *CAF1-55*. The coincidence of both events in the same lineage might suggest that they are related. In fact, even after *CAF1-55DUP* accumulated some adaptive changes, *CAF1-55* and *CAF1-55DUP* could have competed to be incorporated in protein complexes. Thus, it can be envisaged that the changes introduced in *E(Z)*, *ESC* and *SU(Z)12* could have prevented the misincorporation of *CAF1-55DUP* in PRC2 complexes or, alternatively, that they could have allowed the incorporation of *CAF1-55DUP* in PRC2 complexes if *CAF1-55DUP* had a Polycomb-related function.

Caf1-55 enlarges the number of PcG genes duplicated in *Drosophila*. PcG genes code important epigenetic regulators and are mainly single copy genes in the *Drosophila* genus. In fact, PcG proteins form repressive complexes and thus the duplication of PcG genes may hinder maintenance of the proper stoichiometry between the interacting subunits of a protein complex. However, three additional PcG gene duplications have been reported in at least some *Drosophila* species: *Pho/Phol* (Pleiohomeotic and Pleiohomeotic like), *Esc/Escl* (Extra sexcombs and Extra sexcombs like) and *Ph-p/Ph-d* (Polyhomeotic proximal Polyhomeotic distal). A search in the genomes of different Diptera species using *Bombyx mori* as the outgroup was performed to gain a better insight into the duplication events affecting PcG genes and to date the detected duplications (Fig. 4). *Pho* and *Phol* are the most ancient duplicates (~150 Mya) as both paralogs are present in all Diptera analyzed except the two members of the Culicidae family. The *Esc/Escl* duplication took place later, after the split of the Psychodidae family (~130 Mya). The duplication *Ph-p/Ph-d* is more recent (~30 Mya) as the presence of both copies is only shared by the species of the Sophophora subgenus, with the exception of *D. willistoni*. Therefore, the *Caf1-55/Caf1-55dup* duplication in the ancestral branch of the obscure group species included in the Sophophora subgenus is the most recent and took place ~18 Mya. The lack of *Caf1-55dup* in *D. melanogaster* likely contributed to the fact that this duplication has remained undetected until now.

The evolutionary fate of the derived duplicate differs in the four duplicated PcG genes as reflected in the maximum likelihood phylogenetic trees inferred from divergence among paralogs (see Supplementary Fig. S5). As stated above, *Caf1-55dup* most likely underwent a neofunctionalization process. According to its phylogenetic tree, a similar process seems to have affected *Phol* after its origin via an RNA-mediated duplication. *Phol*

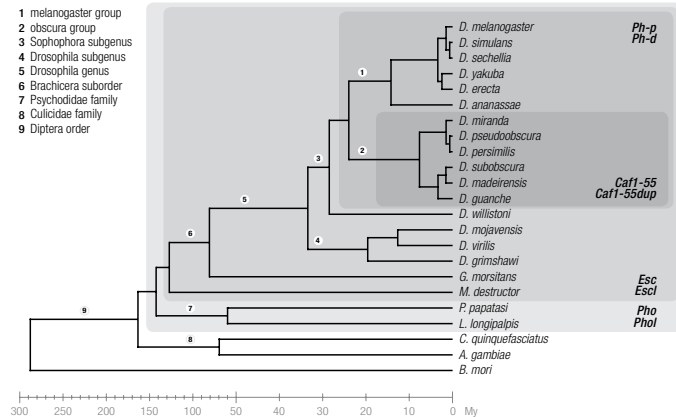


Figure 4. Duplication events of the Polycomb group genes in the Diptera phylogeny. Grey backgrounds group species that share the presence of a particular duplication. The intensity of the shadows ranges from very light grey, which indicates the most ancient duplication (*Pho/Phol*), light grey (*Esc/Escl* duplication), grey (*Ph-p/Ph-d* duplication) and dark grey, which shows the most recent duplication (*Caf1-55/Caf1-55dup*). The bar at the bottom indicates divergence times^{29,55} on two different scales. *G. morsitans* = *Glossina morsitans*, *M. destructor* = *Mayetiola destructor*, *P. papatasi* = *Phlebotomus papatasi*, *L. longipalpis* = *Lutzomyia longipalpis*, *C. quinquefasciatus* = *Culex quinquefasciatus*, *A. gambiae* = *Anopheles gambiae* and *B. mori* = *Bombix mori*.

neofunctionalization is also supported by ChIP on chip experiments that indicate that the binding patterns of PHO and PHOL do not always overlap in the genome⁴², although both proteins bind to the same target DNA sequence *in vitro*. In contrast, *Ph-p* and *Ph-d* are paralogous tandem genes that are known to have evolved under gene conversion⁴³, which is consistent with the clustering of genes and species in the reconstructed phylogenetic tree. The single *Ph* gene present in the *Drosophila* subgenus species and in *D. willistoni* has a structure more similar to *Ph-p* than to *Ph-d*, suggesting that *Ph-p* is the ancestral gene. However, although the coding regions of *Ph-d* and *Ph-p* are very similar due to gene conversion, their regulatory regions are entirely different, suggesting functional divergence at the expression level. Finally, the *Esc/Escl* phylogenetic tree indicates that *Escl* is the ancestral gene and *Esc* the derived duplicate, as already suggested by Ohno *et al.*⁴⁴. However, the divergence between paralogs after the duplication event is much lower than in the case of *Phol* and *Caf1-55dup*. ESC and ESCL, like CAF1-55, are members of the WD40 protein family and both can be incorporated alternatively in the PRC2 Polycomb complex. Nevertheless, ESC is present at high levels during embryogenesis, and ESC-containing PRC2 complexes are critical during early development, whereas the peak abundance of ESCL is found during postembryonic stages⁴⁵. Therefore, the differential expression of the two paralogs during development would suggest a neofunctionalization process mainly at the expression level.

The *Esc/Escl*, *Pho/Phol* and *Ph-d/Ph-p* paralogs are therefore not strictly redundant, although they code Polycomb proteins with related functions. The results obtained by WDSP indicate that CAF1-55DUP is a member of the WD40 protein family, which suggests that it has retained its function as a scaffold to facilitate the interaction between proteins. However, WD40 is one of the most widespread protein families in eukaryotic organisms and WD40 domains are among the most versatile interactors⁹. Therefore, it is not clear whether CAF1-55DUP has a PcG-related function. In fact, although the residues implicated in the formation of hydrogen bonds are highly conserved between CAF1-55 and CAF1-55DUP, the residues that confer the capacity for and specificity of protein–protein interactions are more divergent. The decoupled conservation of these two kinds of residues could indicate that the general function of WD40 is maintained in both paralogs but that they probably interact with different proteins, supporting the neofunctionalization hypothesis.

In addition, the RNA-seq data of *D. pseudoobscura*⁴⁶ available in FlyBase show that *Caf1-55* and *Caf1-55dup* differ in their expression profiles. *Caf1-55* is ubiquitously expressed: at moderate levels in carcass and head, and at high and very high levels in testis and ovary, respectively. In contrast, the expression levels of *Caf1-55dup* are much lower, being moderate only in the testis and ovary (see Supplementary Fig. S6). Functional experiments would be required to characterize the function of *Caf1-55dup* and to explain its differential expression profile.

In summary, the gene encoding the multifunctional and highly conserved protein CAF1-55 is duplicated in the *Drosophila* species of the obscura group. This duplication took place ~18 Mya and enlarges the number of PcG-duplicated genes. The duplicates have suffered clear neofunctionalization, with the action of strong purifying selection on the ancestral copy and of positive selection on the new copy. Positive selection has also acted in a more recent timescale during the divergence of *D. subobscura* and *D. guanche*, as reflected in the *D. subobscura* intraspecific analysis. However, CAF1-55DUP has retained the functional domains of CAF1-55 in all obscura

species, suggesting that it is also a member of the WD40 protein family. Given that proteins of this family are among the most versatile interactors, it is not clear whether CAF1-55DUP has a PcG-related function.

Material and Methods

Fly stocks and sequencing. The *chcu* strain of *D. subobscura* and highly inbred lines of *D. madeirensis* and *D. guanche* were used to sequence *Caf1-55dup* in these species. Nucleotide polymorphism in *Caf1-55* and *Caf1-55dup* was analyzed in highly inbred lines of *D. subobscura* obtained as described in Pratdesaba *et al.*³¹ after sampling a natural population of the species in the Observatori Fabra (Barcelona, Catalonia, Spain). Lines to be studied were chosen by taking into account the chromosomal location of both genes, as inferred by *in situ* hybridization on polytene chromosomes using biotinylated probes⁴⁷. Thus, the selected lines that had either the O_{st} or O_{3+4+8} arrangements, although differing by three overlapping inversions, are homokaryotypic for the proximal half of the O chromosome where *Caf1-55* and *Caf1-55dup* map. Therefore, no effect of the extensive *D. subobscura* inversion polymorphism is expected on the level and pattern of nucleotide variation detected in *Caf1-55* and *Caf1-55dup*. Genomic DNA of these fly stocks, as well as of *D. pseudoobscura*, *D. persimilis* and *D. miranda*, was available in our laboratory.

The *Caf1-55* and *Caf1-55dup* genes were PCR amplified with primers designed according to the *D. pseudoobscura* sequence using the OLIGO program⁴⁸. Amplicons were purified with Multiscreen plates (Millipore) and both strands were completely sequenced with the ABI Prism BigDye Terminators 3.0 Cycle kit (Applied Biosystems) using internal primers. Sequencing reactions were run on an ABI PRISM 3700 sequencer. Partial sequences were assembled using the SEQMAN program of the LASERGENE package⁴⁹. Total RNA of the obscure species was extracted with the RNeasyTM Mini Kit (Qiagen) and then the cDNA was synthesized using the SuperScriptTM III Reverse Transcriptase Kit (Thermo Fisher Scientific), following in both cases the manufacturers' instructions. Subsequently, *Caf1-55dup* cDNA was PCR amplified and sequenced as explained above. The sequences of the primers used in the PCR amplification and sequencing, as well as the PCR conditions, are available in the electronic supplementary material, Table S1.

Divergence analysis. For the species sequenced by the *Drosophila* 12 Genomes Consortium¹⁵ and *D. miranda*, the sequences of *Caf1-55* and its paralogs were retrieved from FlyBase after BLASTN searches using the default parameters. Sequence similarity and synteny conservation with flanking genes were analyzed to distinguish between ancestral and derived gene copies. The *Caf1-55* sequences of *D. subobscura*, *D. madeirensis* and *D. guanche* were retrieved from the EMBL Nucleotide Sequence Database (accession numbers LN864767-69) and those of *Caf1-55dup* were determined in the present study. Orthologous and paralogous sequences were multiply aligned using the MUSCLE program⁵⁰ implemented in the MEGA6 software package⁵¹ according to the alignment of the predicted proteins.

Amino acid conservation along the multiple alignment was inferred for each gene independently using the Clustal X program⁵², which assigns a conservation score for each position of the alignment based on the mean of the distances between codons (according to weight matrix BLOSUM62) and normalized by the percentage of sequences without gaps at this position. The maximum likelihood approach implemented in MEGA6 was used to infer the branch lengths of the accepted phylogenetic tree of the studied species based on nucleotide divergence according to the GTR (general time reversible) model.

The PAML v4 package⁵⁰ was used to compare alternative evolutionary branch models that differ by assumptions concerning the ω estimates ($\omega = d_N/d_S$, where d_N corresponds to nonsynonymous and d_S to synonymous divergence). The M0 model assumes a single ω estimate for all branches, the 3R model assumes three different ω estimates, each one for a different set of branches, and the FR model assumes a different ω for each branch. The branch-site test of positive selection (test 2 in Zhang *et al.*⁵³) implemented in the same package was performed to detect the putative presence of codons under positive selection ($\omega > 1$) in particular branches predefined as foreground branches. The Bayes Empirical Bayes (BEB) method⁵³ also implemented in PAML was used to identify the sites with a high posterior probability of having evolved under positive selection in these branches ($\omega > 1$). In addition, the data set was analyzed by three methods implemented in the HYPHY software package⁵⁴. First, the aBSREL branch-site random effects likelihood method²¹ was used to detect evidence of positive selection in particular branches. This method, in contrast to the branch-site methods implemented in PAML, does not require to predefine foreground branches in the phylogeny. Second, the RELAX approach²² was used to further confirm that increases in ω estimates were indeed due to positive selection and not to relaxed selection. Finally, the BUSTED approach²⁴ was applied to identify the codon sites under positive selection in foreground branches.

The software WDSP^{25,26} was used to determine whether the protein CAF1-55DUP retains the WD40 repeats structure present in CAF1-55. WDSP infers the secondary structure of a given protein, identifies WD40 repeats and estimates a score for each detected WD40 repeat. The tested protein is considered a member of the WD40 family when it presents more than six repeats and the average score of these repeats is greater than 48.

Caf1-55 is not the only duplicated Polycomb gene in *Drosophila*. In fact, other PcG genes are known to be duplicated in *D. melanogaster*: *Ph-p/Ph-d*, *Esc/Escl* and *Pho/Phol*. The OrthoDB hierarchical catalog of orthologous genes (<http://orthodb.org>) was used to infer the presence or absence of these genes in other non-*Drosophila* insect species. The phylogenetic relationships of the insect species in which both paralogs are present and the timing of divergence reported in Misof *et al.*⁵⁵ were used in this analysis.

Estimates of the duplication events of *Caf1-55* were inferred using the BEAST 2 (Bayesian Evolutionary Analysis by Sampling Trees) software platform²⁷. The analysis was performed according to a lognormal relaxed clock and the GTR substitution model. The divergence dates of the *Drosophila* species based on the mutation rate²⁸ were used as calibration points. The MCMC analysis was run with a chain length of 100 million steps, sampling every 10 000 steps.

Nucleotide polymorphism analysis. The assembled sequences of each *D. subobscura* line were aligned using the MUSCLE program⁵⁰. Levels of nucleotide polymorphism were estimated by standard parameters, such as the number of polymorphic sites (S) and nucleotide diversity (π). In coding regions, π was estimated independently for synonymous (π_s) and nonsynonymous (π_n) variation. The pattern of variation was analyzed using the Tajima's D³⁰ statistic. The HKA³² and MK³³ tests were performed to detect a putative decoupling of polymorphism and divergence levels either at silent (noncoding and synonymous) sites between the two genes (HKA test) or between synonymous and nonsynonymous sites of the same gene (MK test). The DnaSP v5 program⁵⁶ was used to perform most of the polymorphism analyses and the HKA program⁵⁷ to perform this neutrality test.

References

- Czermin, B. *et al.* Drosophila enhancer of Zeste/ESC complexes have a histone H3 methyltransferase activity that marks chromosomal Polycomb sites. *Cell* **111**, 185–196 (2002).
- Müller, J. *et al.* Histone methyltransferase activity of a Drosophila Polycomb group repressor complex. *Cell* **111**, 197–208 (2002).
- Tyler, J. K., Bulger, M., Kamakaka, R. T., Kobayashi, R. & Kadonaga, J. T. The p55 subunit of Drosophila chromatin assembly factor 1 is homologous to a histone deacetylase-associated protein. *Mol. Cell. Biol.* **16**, 6149–6159 (1996).
- Tsukiyama, T. & Wu, C. Purification and properties of an ATP-dependent nucleosome remodeling factor. *Cell* **83**, 1011–1020 (1995).
- Marhold, J., Brehm, A. & Kramer, K. The Drosophila methyl-DNA binding protein MBD2/3 interacts with the NuRD complex via p55 and M1-2. *BMC Mol. Biol.* **5**, 20 (2004).
- Beall, E. L. *et al.* Role for a Drosophila Myb-containing protein complex in site-specific DNA replication. *Nature* **420**, 833–7 (2002).
- Korenjak, M. *et al.* Native E2F/RBF complexes contain Myb-interacting proteins and repress transcription of developmentally controlled E2F target genes. *Cell* **119**, 181–93 (2004).
- Nowak, A. J. *et al.* Chromatin-modifying complex component Nurf55/p55 associates with histones H3 and H4 and polycomb repressive complex 2 subunit Su(z)12 through partially overlapping binding sites. *J. Biol. Chem.* **286**, 23388–23396 (2011).
- Stirnemann, C. U., Petsalaki, E., Russell, R. B. & Müller, C. W. WD40 proteins propel cellular networks. *Trends Biochem. Sci.* **35**, 565–74 (2010).
- Chatr-Aryamontri, A. *et al.* The BioGRID interaction database: 2015 update. *Nucleic Acids Res.* **43**, D470–8 (2015).
- Wen, P., Quan, Z. & Xi, R. The biological function of the WD40 repeat-containing protein p55/Caf1 in Drosophila. *Dev. Dyn.* **241**, 455–64 (2012).
- Anderson, A. E. *et al.* The enhancer of trithorax and polycomb gene Caf1/p55 is essential for cell survival and patterning in Drosophila development. *Development* **138**, 1957–1966 (2011).
- Hahn, M. W. & Kern, A. D. Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks. *Mol. Biol. Evol.* **22**, 803–6 (2005).
- Calvo-Martin, J. M., Librado, P., Aguadé, M., Papaceit, M. & Segarra, C. Adaptive selection and coevolution at the proteins of the Polycomb repressive complexes in Drosophila. *Heredity (Edinb.)* **116**, 213–223 (2015).
- Clark, A. G. *et al.* Evolution of genes and genomes on the Drosophila phylogeny. *Nature* **450**, 203–18 (2007).
- Assis, R. & Bachtrog, D. Neofunctionalization of young duplicate genes in Drosophila. *Proc. Natl. Acad. Sci. USA* **110**, 17409–14 (2013).
- Papp, B., Pál, C. & Hurst, L. D. Dosage sensitivity and the evolution of gene families in yeast. *Nature* **424**, 194–7 (2003).
- Zhou, Q. & Bachtrog, D. Sex-Specific Adaptation Drives Early Sex Chromosome Evolution in Drosophila. *Science* (80-.) **337**, (2012).
- Pegueroles, C., Araúz, P. A., Pascual, M. & Mestres, F. A recombination survey using microsatellites: the O chromosome of Drosophila subobscura. *Genetica* **138**, 795–804 (2010).
- Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586–91 (2007).
- Smith, M. D. *et al.* Less is more: an adaptive branch-site random effects model for efficient detection of episodic diversifying selection. *Mol. Biol. Evol.* **32**, 1342–53 (2015).
- Wertheim, J. O., Murrell, B., Smith, M. D., Kosakovsky Pond, S. L. & Scheffler, K. RELAX: Detecting Relaxed Selection in a Phylogenetic Framework. *Mol. Biol. Evol.* **32**, 820–832 (2015).
- Yang, Z., Wong, W. S. W. & Nielsen, R. Bayes empirical Bayes inference of amino acid sites under positive selection. *Mol. Biol. Evol.* **22**, 1107–1118 (2005).
- Murrell, B. *et al.* Gene-wide identification of episodic selection. *Mol. Biol. Evol.* **32**, 1365–71 (2015).
- Wang, Y. *et al.* WDSPdb: a database for WD40-repeat proteins. *Nucleic Acids Res.* **43**, D339–44 (2015).
- Wang, Y., Jiang, F., Zhuo, Z., Wu, X.-H. & Wu, Y.-D. A method for WD40 repeat detection and secondary structure prediction. *PLoS One* **8**, e65705 (2013).
- Bouckaert, R. *et al.* BEAST 2: a software platform for Bayesian evolutionary analysis. *PLoS Comput. Biol.* **10**, e1003537 (2014).
- Obbard, D. J. *et al.* Estimating divergence dates and substitution rates in the Drosophila phylogeny. *Mol. Biol. Evol.* **29**, 3459–73 (2012).
- Hey, J. & Nielsen, R. Multilocus methods for estimating population sizes, migration rates and divergence time, with applications to the divergence of Drosophila pseudoobscura and D. persimilis. *Genetics* **167**, 747–60 (2004).
- Tajima, F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**, 585–595 (1989).
- Pratdesaba, R., Segarra, C. & Aguadé, M. Inferring the demographic history of Drosophila subobscura from nucleotide variation at regions not affected by chromosomal inversions. *Mol. Ecol.* **24**, 1729–41 (2015).
- Hudson, R. R., Kreitman, M. & Aguadé, M. A test of neutral molecular evolution based on nucleotide data. *Genetics* **116**, 153–9 (1987).
- McDonald, J. H. & Kreitman, M. Adaptive protein evolution at the Adh locus in Drosophila. *Nature* **351**, 652–4 (1991).
- Fay, J. C., Wyckoff, G. J. & Wu, C.-I. Positive and Negative Selection on the Human Genome. *Genetics* **158**, 1227–1234 (2001).
- Eyre-Walker, A. Changing effective population size and the McDonald-Kreitman test. *Genetics* **162**, 2017–24 (2002).
- Parsch, J., Zhang, Z. & Baines, J. F. The influence of demography and weak selection on the McDonald-Kreitman test: an empirical study in Drosophila. *Mol. Biol. Evol.* **26**, 691–8 (2009).
- Innan, H. & Kondrashov, F. The evolution of gene duplications: classifying and distinguishing between models. *Nat. Rev. Genet.* **11**, 97–108 (2010).
- Conant, G. C. & Wagner, A. Asymmetric sequence divergence of duplicate genes. *Genome Res.* **13**, 2052–8 (2003).
- Cusack, B. P. & Wolfe, K. H. Not born equal: increased rate asymmetry in relocated and retrotransposed rodent gene duplicates. *Mol. Biol. Evol.* **24**, 679–86 (2007).
- Han, M. V., Demuth, J. P., McGrath, C. L., Casola, C. & Hahn, M. W. Adaptive evolution of young gene duplicates in mammals. *Genome Res.* **19**, 859–67 (2009).
- Jun, J., Ryrkin, P., Hemphill, E. & Nelson, C. Duplication mechanism and disruptions in flanking regions determine the fate of mammalian gene duplicates. *J. Comput. Biol.* **16**, 1253–66 (2009).
- Schüttengruber, B. *et al.* Functional anatomy of polycomb and trithorax chromatin landscapes in Drosophila embryos. *PLoS Biol.* **7**, (2009).

43. Beisswanger, S. & Stephan, W. Evidence that strong positive selection drives neofunctionalization in the tandemly duplicated polyhomeotic genes in *Drosophila*. *Proc. Natl. Acad. Sci. USA* **105**, 5447–5452 (2008).
44. Ohno, K., McCabe, D., Czermin, B., Imhof, A. & Pirrotta, V. ESC, E2F and their roles in Polycomb Group mechanisms. *Mech. Dev.* **125**, 527–541 (2008).
45. Wang, L. *et al.* Alternative ESC and ESC-like subunits of a polycomb group histone methyltransferase complex are differentially deployed during *Drosophila* development. *Mol. Cell. Biol.* **26**, 2637–2647 (2006).
46. Zhang, Y., Sturgill, D., Parisi, M., Kumar, S. & Oliver, B. Constraint and turnover in sex-biased gene expression in the genus *Drosophila*. *Nature* **450**, 233–237 (2007).
47. Segarra, C. & Aguadé, M. Molecular organization of the X chromosome in different species of the obscure group of *Drosophila*. *Genetics* **130**, 513–521 (1992).
48. Rychlik, W. In *PCR Primer Design. Methods in Molecular Biology Vol. 402* (ed. Yuryev, A.) 35–59 (Humana Press Inc, 2007).
49. Burland, T. G. DNASTAR's Lasergene sequence analysis software. *Methods Mol. Biol.* **132**, 71–91 (2000).
50. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–7 (2004).
51. Tamura, K., Stecher, G., Peterson, D., Filipski, A. & Kumar, S. MEGA6: Molecular evolutionary genetics analysis version 6.0. *Mol. Biol. Evol.* **30**, 2725–2729 (2013).
52. Larkin, M. A. *et al.* Clustal W and Clustal X version 2.0. *Bioinformatics* **23**, 2947–8 (2007).
53. Zhang, J., Nielsen, R. & Yang, Z. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. *Mol. Biol. Evol.* **22**, 2472–9 (2005).
54. Kosakovsky Pond, S. L., Frost, S. D. W. & Muse, S. V. HyPhy: hypothesis testing using phylogenies. *Bioinformatics* **21**, 676–9 (2005).
55. Misof, B. *et al.* Phylogenomics resolves the timing and pattern of insect evolution. *Science* (80-.). **346**, 763–767 (2014).
56. Librado, P. & Rozas, J. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics* **25**, 1451–2 (2009).
57. Hey, J. *HKA software* (2010).

Acknowledgements

We thank Montserrat Aguadé for critical comments on the manuscript and Joel O Wertheim for his help and advice in the analysis with the RELAX software. This work was supported by a predoctoral fellowship from the Agència de Gestió d'Ajuts Universitaris i de Recerca (AGAUR), Generalitat de Catalunya, Catalonia, Spain, to JMC-M; and grants BFU2012–35168 and BFU2015–63732 from the Ministerio de Economía y Competitividad, Spain, and 2009SGR-1287 and 2014SGR10555 from the Comissió Interdepartamental de Recerca i Innovació Tecnològica, Catalonia, Spain, to M Aguadé.

Author Contributions

J.M.C.-M. participated in the design of the study, carried out the molecular lab work and data analysis, and helped draft the manuscript. M.P. and C.S. conceived, designed and coordinated the study, and drafted the manuscript. All authors gave final approval for publication.

Additional Information

Accession codes: The newly reported sequences are deposited in the EMBL/GenBank Data Libraries under accession numbers: LT600471 to LT600503

Supplementary information accompanies this paper at <http://www.nature.com/srep>

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Calvo-Martín, J. M. *et al.* Evidence of neofunctionalization after the duplication of the highly conserved Polycomb group gene *Caf1-55* in the obscure group of *Drosophila*. *Sci. Rep.* **7**, 40536; doi: 10.1038/srep40536 (2017).

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



This work is licensed under a Creative Commons Attribution 4.0 International License. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>

© The Author(s) 2017

Evidence of neofunctionalization after the duplication of the highly conserved Polycomb group gene *Caf1-55* in the obscura group of *Drosophila*

Calvo-Martín JM, Papaceit M, Segarra C

2017. *Scientific Reports* 7:40536. doi: 10.1038/srep40536

Supplementary Material

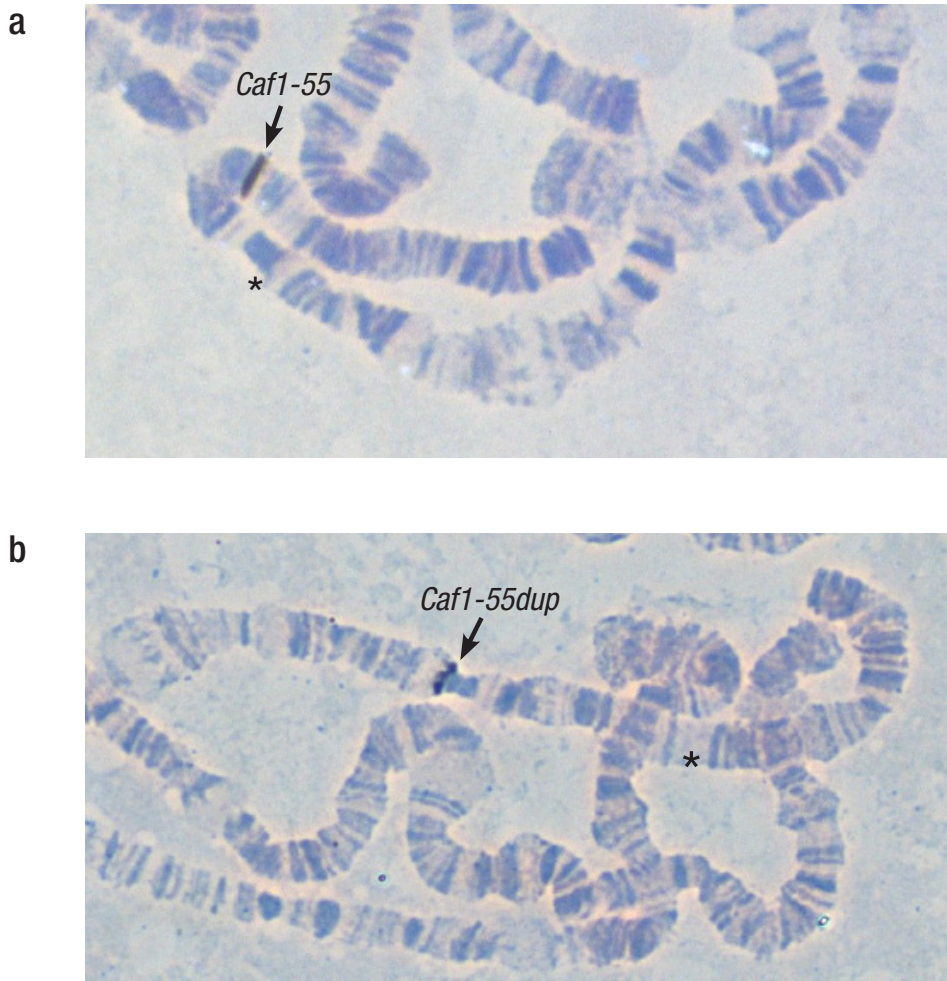


Figure S1. Location of *Caf1-55* (A) and *Caf1-55dup* (B) by in situ hybridization on *D. subobscura* polytene chromosomes. Arrows indicate the position of each gene according to the hybridization site of the biotinylated probes and asterisks the position of its paralog. *Caf1-55* and *Caf1-55dup* are located in the cytogenetic bands 79C and 80D, respectively.

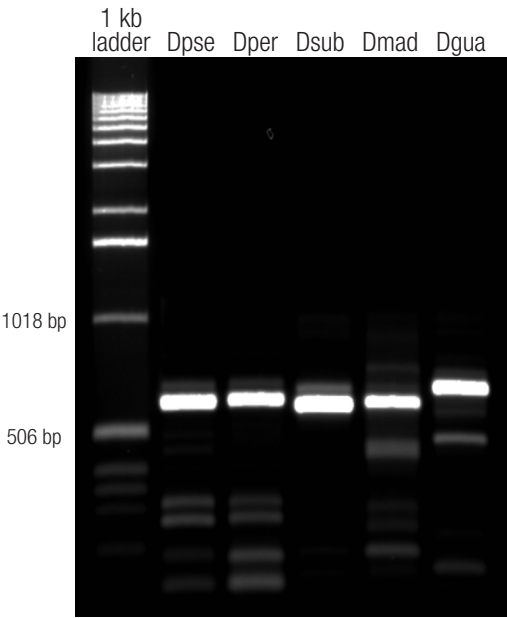


Figure S2. PCR amplification of the second and third exons of *Caf1-55dup* using the cDNA of five *Drosophila* species of the *obscura* group as template. Fragment sizes and posterior sequencing confirm that the recovered and amplified cDNA proceeds from processed mRNA. Species names as in Figure 2.

Carl-55

	5	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	
OF01	C	G	C	G	T	A	A	C	C	A	T	G	C	T	G	A	C	T	T	G	A	T	T	G	A	T
OF07	A	A	G	C	T	T	A	A	C	C	A	T	G	C	T	G	A	C	T	T	G	A	T	T	G	A
OF14	A	A	G	C	T	T	A	A	C	C	A	T	G	C	T	G	A	C	T	T	G	A	T	T	G	A
OF16	T	A	G	T	T	A	A	C	C	A	T	G	C	T	G	A	C	T	T	G	A	T	T	G	A	
OF18	T	T	A	G	C	T	T	A	A	C	C	A	T	G	C	T	G	A	C	T	T	G	A	T	T	
OF19	T	T	A	G	C	T	T	A	A	C	C	A	T	G	C	T	G	A	C	T	T	G	A	T	T	
OF20	T	T	A	G	C	T	T	A	A	C	C	A	T	G	C	T	G	A	C	T	T	G	A	T	T	
OF21	T	T	A	G	C	T	T	A	A	C	C	A	T	G	C	T	G	A	C	T	T	G	A	T	T	
OF22	T	T	A	G	C	T	T	A	A	C	C	A	T	G	C	T	G	A	C	T	T	G	A	T	T	
OF23	T	T	A	G	C	T	T	A	A	C	C	A	T	G	C	T	G	A	C	T	T	G	A	T	T	
OF24	T	T	A	G	C	T	T	A	A	C	C	A	T	G	C	T	G	A	C	T	T	G	A	T	T	
OF25	T	T	A	G	C	T	T	A	A	C	C	A	T	G	C	T	G	A	C	T	T	G	A	T	T	
OF26	T	T	A	G	C	T	T	A	A	C	C	A	T	G	C	T	G	A	C	T	T	G	A	T	T	
OF27	T	T	A	G	C	T	T	A	A	C	C	A	T	G	C	T	G	A	C	T	T	G	A	T	T	
OF28	T	T	A	G	C	T	T	A	A	C	C	A	T	G	C	T	G	A	C	T	T	G	A	T	T	
OF29	T	T	A	G	C	T	T	A	A	C	C	A	T	G	C	T	G	A	C	T	T	G	A	T	T	
OF30	T	T	A	G	C	T	T	A	A	C	C	A	T	G	C	T	G	A	C	T	T	G	A	T	T	
OF31	T	T	A	G	C	T	T	A	A	C	C	A	T	G	C	T	G	A	C	T	T	G	A	T	T	
OF32	T	T	A	G	C	T	T	A	A	C	C	A	T	G	C	T	G	A	C	T	T	G	A	T	T	
OF33	T	T	A	G	C	T	T	A	A	C	C	A	T	G	C	T	G	A	C	T	T	G	A	T	T	
OF34	T	T	A	G	C	T	T	A	A	C	C	A	T	G	C	T	G	A	C	T	T	G	A	T	T	
OF35	T	T	A	G	C	T	T	A	A	C	C	A	T	G	C	T	G	A	C	T	T	G	A	T	T	
OF36	T	T	A	G	C	T	T	A	A	C	C	A	T	G	C	T	G	A	C	T	T	G	A	T	T	
OF37	T	T	A	G	C	T	T	A	A	C	C	A	T	G	C	T	G	A	C	T	T	G	A	T	T	
OF38	T	T	A	G	C	T	T	A	A	C	C	A	T	G	C	T	G	A	C	T	T	G	A	T	T	
OF39	T	T	A	G	C	T	T	A	A	C	C	A	T	G	C	T	G	A	C	T	T	G	A	T	T	
OF40	T	T	A	G	C	T	T	A	A	C	C	A	T	G	C	T	G	A	C	T	T	G	A	T	T	
OF41	T	T	A	G	C	T	T	A	A	C	C	A	T	G	C	T	G	A	C	T	T	G	A	T	T	
OF42	T	T	A	G	C	T	T	A	A	C	C	A	T	G	C	T	G	A	C	T	T	G	A	T	T	
OF43	T	T	A	G	C	T	T	A	A	C	C	A	T	G	C	T	G	A	C	T	T	G	A	T	T	
OF44	T	T	A	G	C	T	T	A	A	C	C	A	T	G	C	T	G	A	C	T	T	G	A	T	T	
OF45	T	T	A	G	C	T	T	A	A	C	C	A	T	G	C	T	G	A	C	T	T	G	A	T	T	
OF46	T	T	A	G	C	T	T	A	A	C	C	A	T	G	C	T	G	A	C	T	T	G	A	T	T	
OF47	T	T	A	G	C	T	T	A	A	C	C	A	T	G	C	T	G	A	C	T	T	G	A	T	T	
OF48	T	T	A	G	C	T	T	A	A	C	C	A	T	G	C	T	G	A	C	T	T	G	A	T	T	
OF49	T	T	A	G	C	T	T	A	A	C	C	A	T	G	C	T	G	A	C	T	T	G	A	T	T	
OF50	T	T	A	G	C	T	T	A	A	C	C	A	T	G	C	T	G	A	C	T	T	G	A	T	T	
OF51	T	T	A	G	C	T	T	A	A	C	C	A	T	G	C	T	G	A	C	T	T	G	A	T	T	
OF52	T	T	A	G	C	T	T	A	A	C	C	A	T	G	C	T	G	A	C	T	T	G	A	T	T	
OF53	T	T	A	G	C	T	T	A	A	C	C	A	T	G	C	T	G	A	C	T	T	G	A	T	T	
OF54	T	T	A	G	C	T	T	A	A	C	C	A	T	G	C	T	G	A	C	T	T	G	A	T	T	
OF55	T	T	A	G	C	T	T	A	A	C	C	A	T	G	C	T	G	A	C	T	T	G	A	T	T	
OF56	T	T	A	G	C	T	T	A	A	C	C	A	T	G	C	T	G	A	C	T	T	G	A	T	T	
OF57	T	T	A	G	C	T	T	A	A	C	C	A	T	G	C	T	G	A	C	T	T	G	A	T	T	
OF58	T	T	A	G	C	T	T	A	A	C	C	A	T	G	C	T	G	A	C	T	T	G	A	T	T	
OF59	T	T	A	G	C	T	T	A	A	C	C	A	T	G	C	T	G	A	C	T	T	G	A	T	T	
OF60	T	T	A	G	C	T	T	A	A	C	C	A	T	G	C	T	G	A	C	T	T	G	A	T	T	
OF61	T	T	A	G	C	T	T	A	A	C	C	A	T	G	C	T	G	A	C	T	T	G	A	T	T	
OF62	T	T	A	G	C	T	T	A	A	C	C	A	T	G	C	T	G	A	C	T	T	G	A	T	T	
OF63	T	T	A	G	C	T	T	A	A	C	C	A	T	G	C	T	G	A	C	T	T	G	A	T	T	
OF64	T	T	A	G	C	T	T	A	A	C	C	A	T	G	C	T	G	A	C	T	T	G	A	T	T	
OF65	T	T	A	G	C	T	T	A	A	C	C	A	T	G	C	T	G	A	C	T	T	G	A	T	T	
OF66	T	T	A	G	C	T	T	A	A	C	C	A	T	G	C	T	G	A	C	T	T	G	A	T	T	
OF67	T	T	A	G	C	T	T	A	A	C	C	A	T	G	C	T	G	A	C	T	T	G	A	T	T	
OF68	T	T	A	G	C	T	T	A	A	C	C	A	T	G	C	T	G	A	C	T	T	G	A	T	T	
OF69	T	T	A	G	C	T	T	A	A	C	C	A	T	G	C	T	G	A	C	T	T	G	A	T	T	
OF70	T	T	A	G	C	T	T	A	A	C	C	A	T	G	C	T	G	A	C	T	T	G	A	T	T	
OF71	T	T	A	G	C	T	T	A	A	C	C	A	T	G	C	T	G	A	C	T	T	G	A	T	T	
OF72	T	T	A	G	C	T	T	A	A	C	C	A	T	G	C	T	G	A	C	T	T	G	A	T	T	
OF73	T	T	A	G	C	T	T	A	A	C	C	A	T	G	C	T	G	A	C	T	T	G	A	T	T	
OF74	T	T	A	G	C	T	T	A	A	C	C	A	T	G	C	T	G	A	C	T	T	G	A	T	T	
OF75	T	T	A	G	C	T	T	A	A	C	C	A	T	G	C	T	G	A	C	T	T	G	A	T	T	
OF76	T	T	A	G	C	T	T	A	A	C	C	A	T	G	C	T	G	A	C	T	T	G	A	T	T	
OF77	T	T	A	G	C	T	T	A	A	C	C	A	T	G	C	T	G	A	C	T	T	G	A	T	T	
OF78	T	T	A	G	C	T	T	A	A	C	C	A	T	G	C	T	G	A	C	T	T	G	A	T	T	
OF79	T	T	A	G	C	T	T	A	A	C	C	A	T	G	C	T	G	A	C	T	T	G	A	T	T	
OF80	T	T	A	G	C	T	T	A	A	C	C	A	T	G	C	T	G	A	C	T	T	G	A	T	T	
OF81	T	T	A	G	C	T	T	A	A	C	C	A	T	G	C	T	G	A	C	T	T	G	A	T	T	
OF82	T	T	A	G	C	T	T	A	A	C	C	A	T	G	C	T	G	A	C	T	T	G	A	T	T	
OF83	T	T	A	G	C	T	T	A	A	C	C	A	T	G	C	T	G	A	C	T	T	G	A	T	T	
OF84	T	T	A	G	C	T	T	A	A	C	C	A	T	G	C	T	G	A	C	T	T	G	A	T	T	
OF85	T	T	A	G	C	T	T	A	A	C	C	A	T	G	C	T	G	A	C	T	T	G	A	T	T	
OF86	T	T	A	G	C	T	T	A	A	C	C	A	T	G	C	T	G	A	C	T	T	G	A	T	T	
OF87	T	T	A	G	C	T	T	A	A	C	C	A	T	G	C	T	G	A	C	T	T	G	A	T	T	
OF88	T	T	A	G	C	T	T	A	A	C	C	A	T	G	C	T	G	A	C	T	T	G	A	T	T	
OF89	T	T	A	G	C	T	T	A	A	C	C	A	T	G	C	T	G	A	C	T	T	G	A	T	T	
OF90	T	T	A	G	C	T	T	A	A	C	C	A	T	G	C	T	G	A	C	T	T	G	A	T	T	
OF91	T	T	A	G	C	T	T	A	A	C	C	A	T	G	C	T	G	A	C	T	T	G	A	T	T	
OF92	T	T	A	G	C	T	T	A	A	C	C	A	T	G	C	T	G	A	C	T	T	G	A	T	T	
OF93	T	T	A	G	C	T	T	A	A	C	C	A	T	G	C	T	G	A	C	T	T	G	A	T	T	
OF94	T	T	A	G	C	T	T	A	A	C	C	A	T	G	C	T	G	A	C	T	T	G	A	T	T	
OF95	T	T	A	G	C	T	T	A	A	C	C	A	T	G	C	T	G	A	C	T	T	G	A	T	T	
OF96	T	T	A	G	C	T	T	A	A	C	C	A	T	G	C	T	G	A	C	T	T	G	A	T	T	
OF97	T	T	A	G	C	T	T	A	A	C	C	A	T	G	C	T	G	A	C	T	T	G	A	T	T	
OF98	T	T	A	G	C	T	T	A	A	C	C	A	T	G	C	T	G	A	C	T	T	G	A	T	T	
OF99	T	T	A	G	C	T	T	A	A	C	C	A	T	G	C	T	G	A	C	T	T	G	A	T	T	
OF100	T	T	A	G	C	T	T	A	A	C	C	A	T	G	C	T	G	A	C	T	T	G	A	T	T	

4-55dup

Figure S4. *Caf1-55* and *Caf1-55dup* nucleotide polymorphic sites in 14 and 16 lines of *D. subobscura*, respectively. The first column indicates the name of each sequenced *D. subobscura* line. Numbers show nucleotide sites along the multiple alignment. Dots indicate the same nucleotide as in the QF01 sequence (reference). Asterisks indicate alignment gaps. Polymorphic sites marked with an asterisk correspond to nonsynonymous polymorphisms and the site marked with a cross in *Caf1-55dup* represents the gain of a stop codon. Information for the polymorphic sites in the *D. madeirensis* and *D. guanche* sequences used as outgroups is also shown. ln_1 = insertion of n nucleotides; $(A)_n$ = n repeats of adenine, ? = not sequenced. Bars above both tables indicate if a certain site belongs to 5' = 5' flanking region; 11 = first intron; E2 = second intron; E3 = third intron; 3' = 3' flanking region.



Figure S5. Phylogenetic trees according to nucleotide divergence of the *Esc/Escl*, *Pho/Phol* and *Ph-p/Ph-d* paralogs. Branch lengths were obtained with MEGA6 by maximum likelihood fixing the commonly accepted phylogeny of the studied species. The ancestral genes are represented in black and the derived duplicates in red. The scale at the lower right corner of the figure that indicates nucleotide substitutions per site is common for the three trees. Asterisks indicate the nodes where the duplication events occurred. Branches too long relative to the other branches are interrupted by a double dash.

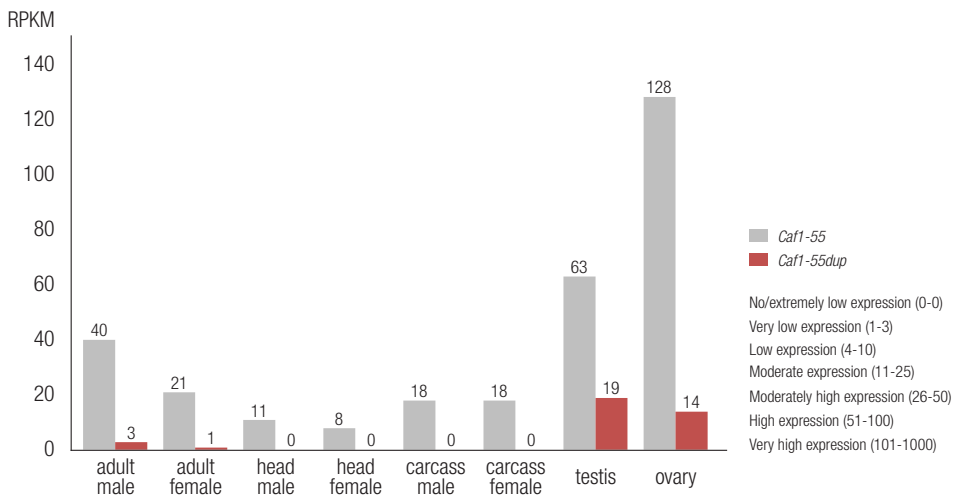


Figure S6. Expression levels of *Caf1-55* and *Caf1-55dup* in *D. pseudoobscura*, according to the RNA-Seq data available in FlyBase. Values correspond to reads per kilobase of exon per million mapped reads (RPKM). The experiments for each tissue are independent. Thus, RPKM values for the whole adult flies do not correspond to the sum of the RPKM values for the separate tissues. Testis or ovary were not included when carcass expression was analyzed. Clasification of the expression levels, from “no expression” to “very high expression”, is given according to FlyBase.

Table S1. Primers used to amplify and sequence the studied genes, as well as PCR conditions.

<i>Caf1-55</i>	Amplification primers		Annealing temperature ^a	Extension time ^b	Sequencing primers	
<i>D. subobscura</i>	C1_F1	5' CAGCCGAGATTCAGTGAGC 3'	56 °C	1'20"	C1_F1	5' CAGCCGAGATTCAGTGAGC 3'
<i>D. madeirensis</i>	C1_R1	5' CCATCTTCGGCATCCTCTG 3'			C1_R1	5' CCATCTTCGGCATCCTCTG 3'
<i>D. guanche</i>						
	C1_F2	5' CGGTGAATGCCAGCCAGAT 3'	56 °C	1'20"	C1_F2	5' CGGTGAATGCCAGCCAGAT 3'
	C1_R2	5' TCAGCACATCGCCCTATCA 3'			C1_R2	5' TCAGCACATCGCCCTATCA 3'
					C1_SR1	5' CCTTCTCGTTGTCATAGTG 3'
					C1_SR2	5' CACAAGCAGATGGTATGAT 3'
					C1_SR3	5' GGAAGAGGCAAGAATGGTC 3'
					C1_SR4	5' CCAGAGCGTGTGTCATTAC 3'
					C1_SR5	5' TGCTGCTGCCCAAGTATG 3'

GL12106						
<i>D. persimilis</i>	C1_F2	5' CGGTGAATGCCAGCCAGAT 3'	54 °C	2'00"	C1_F2	5' CGGTGAATGCCAGCCAGAT 3'
	C1_R3	5' CCACTGCCAGGATCGTTCT 3'			C1_R3	5' CCACTGCCAGGATCGTTCT 3'
					C1_R2	5' TCAGCACATCGCCCTATCA 3'
					C1_SR3	5' GGAAGAGGCAAGAATGGTC 3'

<i>Caf1-55dup</i>						
<i>D. subobscura</i>	C1D_F1	5' GTTTCGCTTGGCATGTCATAC 3'	54 °C	2'10"	C1D_F1	5' GTTTCGCTTGGCATGTCATAC 3'
<i>D. madeirensis</i>	C1D_R1	5' ACACATCTTGACCGTTGACA 3'			C1D_R1	5' ACACATCTTGACCGTTGACA 3'
<i>D. guanche</i>						
					C1D_SF1	5' CTGTACGATGAGATTGTGAC 3'
					C1D_SR1	5' TCGGACACGGAGCACATGAT 3'
					C1D_SF2	5' ATTGCCCTTGTGGGATATG 3'
					C1D_SR2	5' AGCGTTGTGTCCACTGAA 3'
					C1D_SR3	5' ACACCAGGACATCGCTCTT 3'
					C1D_SR4	5' AGCCACTGGACCGTAAAGGAT 3'

<i>D. miranda</i>	C1D_F2	5' AGGGAGCGAAATGTGCAATC 3'	54 °C	40"	C1D_F2	5' AGGGAGCGAAATGTGCAATC 3'
	C1D_R2	5' ATGTAGCGAGCACGGTTCAC 3'			C1D_R2	5' ATGTAGCGAGCACGGTTCAC 3'

<i>Caf1-55dup cDNA</i>						
<i>D. pseudoobscura</i>	C1D_F3	5' GTCATCATCAACCACGAAGG 3'	57 °C	40"	C1D_F3	5' GTCATCATCAACCACGAAGG 3'
<i>D. persimilis</i>	C1D_R3a	5' CATTTGCCACACTTCCATCA 3'			C1D_R3a	5' CATTTGCCACACTTCCATCA 3'
					C1D_SR5a	5' TATCCACAAAGGCACAGT 3'

<i>D. subobscura</i>	C1D_F3	5' GTCATCATCAACCACGAAGG 3'	57 °C	40"	C1D_F3	5' GTCATCATCAACCACGAAGG 3'
<i>D. madeirensis</i>	C1D_R3b	5' TCCATGAGAAGTCGTTGATT 3'			C1D_R3b	5' TCCATGAGAAGTCGTTGATT 3'
<i>D. guanche</i>					C1D_SR5b	5' ATCCCACAAGGCAATGTC 3'

^a Used to synthesize the cDNA

Common PCR conditions	32 cycles				
	Initial denaturation	Denaturation	Annealing	Extension	Final extension
	94 °C 5'00"	96 °C 10"	a 10"	65 °C b	65 °C 5'00"

Artículo 3

Molecular population genetics of the Polycomb genes in *Drosophila subobscura*

Calvo-Martín JM, Papaceit M, Segarra C

Pendiente de publicación

Las proteínas del grupo Polycomb (PcG) son importantes moduladores del estado de la cromatina. Forman complejos proteicos que reprimen la expresión génica mediante la introducción de modificaciones postraduccionales en las histonas. El estudio de la divergencia de las proteínas PcG en *Drosophila* ha revelado señales de coevolución entre ellas, así como la acción de selección adaptativa en el linaje ancestral de las especies del grupo *obscura*, principalmente en las subunidades del complejo Pcl-PRC2. En el presente trabajo se ha estudiado el polimorfismo nucleotídico de los genes PcG en una población natural de *D. subobscura* con el fin de detectar si la selección natural ha modulado también la evolución de estos importantes genes reguladores a una escala de tiempo más reciente. La mayoría de los genes estudiados presentan niveles y patrones de polimorfismo consistentes con las predicciones del modelo neutro estándar, siendo las únicas excepciones los genes *Su(z)12* y *Pho*. Así, el test MK infiere una acumulación de cambios adaptativos en la proteína SU(Z)12 durante la divergencia de *D. subobscura* y *D. guanche*. Por otro lado, el test HKA muestra un déficit de polimorfismo en el gen *Pho*. Sin embargo, la explicación más probable para esta reducida variación no es la acción de la selección adaptativa actuando directamente sobre *Pho*, sino la localización de este gen en el cromosoma puntiforme de *D. subobscura*. De hecho, las características de la variación nucleotídica de *Pho* en esta especie concuerdan con las publicadas previamente en genes localizados en el cromosoma puntiforme de *D. melanogaster* y otras especies de *Drosophila*, caracterizado por su ausencia de recombinación.

Molecular population genetics of the Polycomb genes in *Drosophila subobscura*

Calvo-Martín JM, Papaceit M, Segarra C

ABSTRACT

Polycomb group (PcG) proteins are important regulatory proteins that modulate the chromatin state. They form protein complexes that repress gene expression by the introduction of posttranslational histone modifications. The study of PcG proteins divergence in *Drosophila* revealed signals of coevolution among them and the action of positive selection in the lineage ancestral to the *obscura* group species, mainly in subunits of the Pcl-PRC2 complex. Herein, we have studied the nucleotide polymorphism of PcG genes in a natural population of *D. subobscura* to detect whether natural selection has also modulated the evolution of these important regulatory genes in a more recent time scale. Most genes present a level and pattern of polymorphism consistent with predictions of the neutral model, the exceptions being *Su(z)12* and *Pho*. Thus, the MK test indicates an accumulation of adaptive changes in the SU(Z)12 protein during the divergence of *D. subobscura* and *D. guanche*. In contrast, the HKA test shows a deficit of polymorphism at *Pho*. However, the most likely explanation for this reduced variation is not the action of positive selection acting directly on *Pho*, but the location of this gene in the dot-like chromosome of *D. subobscura*. In fact, the characteristics of nucleotide variation at *Pho* in this species conform to those previously reported in genes located in the nonrecombining dot-like chromosome of *D. melanogaster* and other *Drosophila* species.

INTRODUCTION

The identification of genes under adaptive selection is a major goal of evolutionary genetics. Several methods and statistical tests have been developed to detect the footprint left by the action of positive selection at the molecular level (reviewed in Vitti, Grossman and Sabeti 2013). These different approaches were first applied to candidate genes with a clear adaptive function through the analysis of nucleotide divergence between species, nucleotide polymorphism within species or both. In *Drosophila*, these studies consistently corroborated that adaptive selection has shaped the evolution of genes involved in processes such as immunity (Sackton et al. 2007), reproduction (Wong et al. 2008) and stem cell maintenance (Flores et al. 2015). The richness of these functional gene categories in rapidly evolving genes has been further confirmed in whole genome population studies both in *D. melanogaster* (Langley et al. 2012; Mackay et al. 2012) and *D. simulans* (Begun et al. 2007).

Molecular evolution of proteins and thus of their encoding genes is, however, not solely determined by selection on protein function (Pál, Papp and Lercher 2006). In fact, proteins do not act in isolation, but they are members of complex metabolic, regulatory or interaction networks that control biological processes. The evolution of a protein can be affected by its location in the network, the connectivity with network partners and the evolution of the network as a whole (Cork and Purugganan 2004). Thus, the strength of purifying selection on genes coding for enzymes depends on the protein position either in the metabolic pathway (Eanes 2011) or in the signal transduction pathway (Alvarez-Ponce et al. 2012). In addition, physically interacting proteins tend to evolve co-ordinately. Accordingly, pervasive adaptive evolution has been detected among genes coding for subunits of nuclear pore complexes (Presgraves and Stephan 2007) and centromeric interacting proteins (Beck and Llopart 2015). This concordant evolution also results in coincident increases or decreases of the evolutionary rate over particular branches of a phylogeny, as has been detected among genes encoding either interacting proteins (Clark and Aquadro 2010) or members of protein networks (Findlay et al. 2014).

Polycomb group (PcG) genes code for proteins that form diverse protein repressive complexes involved in the epigenetic silencing of chromatin by the introduction of posttranslational histone modifications (Lanzuolo and Orlando 2012). Evidence of coevolution among the subunits of these complexes has been detected in an interespecific divergence study in *Drosophila*, which also revealed acceleration in the nonsynonymous substitution rate in the lineage ancestral to the *obscura* group species, mainly in genes encoding subunits of the Pcl-PRC2 complex

(Calvo-Martín et al. 2016). Therefore, signatures of adaptive selection have been detected in some PcG genes prior diversification of the *obscura* species. According to these results, it is likely that positive selection might also have modulated PcG genes evolution in a more recent time scale.

Here, we have analyzed by high-quality Sanger sequencing the nucleotide polymorphism of 16 PcG genes that code for the subunits of four Polycomb repressive complexes (PhoRC, Pcl-PRC2, PRC1 and dRAF) in *D. subobscura* as a representative of the *obscura* group species. In contrast to our expectations, no pervasive signatures of recent adaptive evolution were detected in the Polycomb group genes. In fact, evidence of selection was only detected at *Su(z)12*, which encodes a subunit of the Pcl-PRC2 complex. This gene showed an excess of adaptive changes during the divergence of *D. subobscura* and *D. guanche* according to the MK test (McDonald and Kreitman 1991). In addition, another outstanding result is the extremely low level of variation at *Pho* that, however, cannot be associated with a recent selective sweep in this gene in *D. subobscura*, but to its location in the dot-like chromosome. In fact, the dot-like chromosome has a wide footprint of reduced variation in *D. melanogaster* and other species of the *melanogaster* group (Wang et al. 2002, 2004; Arguello et al. 2010), which is consistent with its virtually lack of recombination in these species. Therefore, present results would rather support that the dot-like chromosome of *D. subobscura* also has null or very low recombination rates, which would cause a deficit of variation at *Pho* and linked genes.

MATERIAL AND METHODS

Genes and fly stocks

The sixteen Polycomb group (PcG) genes studied code for subunits of four Polycomb repressive complexes: PhoRC, Pcl-PRC2, PRC1 and dRAF. Genes are: *Pho*, its paralog *Phol* and *Sfmbt* (PhoRC complex); *Caf1-55*, *E(z)*, *Esc*, its paralog *Escl*, *Su(z)12* and *Pcl* (Pcl-PRC2 complex); *Psc*, *Sce*, *Pc*, *Ph-p*, its paralog *Ph-d* and *Scm* (PRC1 complex), and *Kdm2* (dRAF complex that also contains PSC and SCE). The sequences of these genes in *D. subobscura*, *D. madeirensis* and *D. guanche* are available in EMBL/GenBank Data Libraries under accession numbers LN864767 to LN864769 and LN864773 to LN864814. The location of these genes in the polytene chromosomes of the *chcu* strain of *D. subobscura* that is homokaryotypic for known chromosomal arrangements was determined by *in situ* hybridization using biotinylated probes, as described in Papaceit, Segarra and Aguadé (2013).

Highly inbred lines of *D. subobscura* were obtained by sib mating from isofemales lines established after sampling a natural population in the Observatori Fabra (outskirts of Barcelona, Catalonia, Spain). About half of the studied lines were already available in our laboratory and had been previously studied by Pratdesaba, Segarra and Aguadé (2015). *D. subobscura* has a rich chromosomal polymorphism. Therefore, the studied lines for each gene were carefully selected to prevent any effect of chromosomal polymorphism on the level and pattern of nucleotide variation. For instance, *Ph-p* and *Ph-d* are located towards the distal end of the A (=X) chromosome near one of the breakpoints of the A₂ inversion. Consequently, variation at *Ph-p* and *Ph-d* was studied only in A_{st} and A₁ lines that are homokaryotypic for the distal half of the chromosome were the genes map. Supplementary Table S1 indicates the chromosome and chromosomal section where each gene is located and the arrangement for this chromosome of the selected lines.

DNA sequencing and analysis

Genomic DNA of the selected *D. subobscura* lines was purified with the Puregen®Core Kit B (Qiagen). Nucleotide variation at *Caf1-55* in *D. subobscura* has been previously reported (Calvo-Martín, Papaceit and Segarra 2017) and sequences are available in EMBL/GenBank Data Libraries under accession numbers LT600474 to LT600487. The polymerase chain reaction was used to amplify the complete coding region of the additional 15 PcG genes here studied and the obtained PCR products were purified with Multiscreen plates (Millipore). The ABI Prism BigDye Terminators 3.0 Cycle kit (Applied Biosystems) and internal primers were used to sequence completely both strands of the amplicons. Sequencing reactions were run on an ABI PRISM 3700 sequencer at Serveis Científic-Tècnics de la Universitat de Barcelona. PCR conditions, and the sequences of the primers used in the PCR amplification and sequencing are available on request to the authors. Partial sequences of each gene were assembled with the SEQMAN program of the DNASTAR Lasergene package (Burland 2000). Complete sequences were multiply aligned with the MEGA6 software (Tamura et al. 2013).

The DnaSP v5 program (Librado and Rozas 2009) was used to estimate the basic summary statistics that describe nucleotide variation, to measure the extent of codon bias and to perform most neutrality tests. The level and pattern of variation was estimated by the number of polymorphic sites (*S*), nucleotide diversity (π), haplotype diversity (*Hd*), Tajima's *D* statistic (Tajima 1989), the minimum number of recombination events (*R_m*) inferred by the four-gamete test (Hudson and

Kaplan 1985) and the linkage disequilibrium between parsimony informative sites measured by the $|D'|$ parameter (Lewontin 1964). Codon bias was estimated by the effective number of codons (ENC) as proposed by Wright (1990). ENC ranges from 20 when a single codon is used for each amino acid (maximum codon bias) to 64 when all synonymous codons are used equally (no codon bias).

Different neutrality tests were applied to the data to detect signatures of selection. Thus, the Tajima's test (Tajima 1989) based only on polymorphism data and the different Fu and Li's tests (Fu and Li 1993) were used to infer whether the observed pattern of variation conforms to that expected under the neutral model in a population at mutation-drift equilibrium. Fu and Li's tests based on the D^* and F^* statistics can be performed when only polymorphism data in a single species are available, in contrast to those based on the D and F statistics that also require data on an outgroup species. In addition, the HKA test (Hudson, Kreitman and Aguadé 1987) and the MK test (McDonald and Kreitman 1991) were applied to detect a putative decoupling of polymorphism and divergence among genes (HKA test) and between synonymous and nonsynonymous sites within a gene (MK test). The multilocus HKA tests was performed by the HKA program developed by J. Hey and available as a software resource at <https://bio.cst.temple.edu/~hey>. This program conducts coalescence simulations to infer the significance of the observed χ^2 -like test statistic and to perform the maximum cell value test. As *D. subobscura* has a genome wide pattern of variation consistent with an expansion process, significance levels of *Hd*, Tajima's D and Fu and Li's D and F tests statistics were estimated by computer simulations using the software mlcoalsim (Ramos-Onsins and Mitchell-Olds 2007) with the parameters of the expansion model inferred in Pratdesaba, Segarra and Aguadé (2015).

RESULTS

The chromosomal location in *D. subobscura* of the sixteen studied Polycomb group (PcG) genes is shown in Figure 1. Genes map at the expected chromosome according to their cytological location in *D. melanogaster* and the well-established chromosomal homologies between this species and *D. subobscura*. Only the two tandem paralogs *Ph-p* and *Ph-d* are X-linked, all other genes are located in autosomes and *Pho* maps in the dot-like chromosome (Muller's element F) that presents an exceedingly low recombination rate in *D. melanogaster* (Sandler and Szauter 1978). Chromosomal locations were taken into account when selecting



Figure 1. The location of the sixteen studied PcG genes is indicated on the chromosomes of *D. subobscura* (*chcu* strain) by arrowheads. All chromosomes have the standard arrangement except the O chromosome that has the O_{3+4} arrangement. Chromosomes are identified by capital letters (A, J, U, E and O) at the distal end. The dot-like chromosome is also indicated (Dot).

the chromosomal arrangement of the lines sequenced for each gene to prevent any effect of inversions on nucleotide variation (Supplementary Table S1).

Polymorphic sites and indels present in the multiple alignment of each gene are shown in Supplementary Figures S1a to S1p. Summary statistics of nucleotide variation are shown in Table 1. Nucleotide diversity (π) ranges from 0.0009 at *Pho* to 0.0116 at *Escl*. The low level of variation at *Pho* relative to the other PcG gene regions is more evident when nucleotide diversity is normalized by nucleotide divergence (K) with *D. guanche* to correct for differences in the mutation rate across gene regions. The π/K ratio at *Pho* is 0.0260 that is also the lowest estimate and about 3.2 times lower than the next value present at *Pcl* (0.0833). In addition to nucleotide diversity, *Pho* also has the lowest number of haplotypes (10 in 17 lines) and haplotype diversity ($Hd = 0.9$). No recombination events were inferred in the *Pho* region ($R_m = 0$) by the four-gamete test. *Pho* also present a rather low codon bias ($ENC = 54.98$). This estimate is slightly higher (indicating a lower codon bias) than the median ENC value (53.92) for the *D. melanogaster* genes located in the dot-like element (Leung et al. 2010). Tajima's D statistic was negative in all regions (Table 1), which is not surprising given that *D. subobscura* has a genome

wide pattern of variation consistent with an expansion process (Pratdesaba, Segarra and Aguadé 2015). None of the performed Tajima's test was significant ($P > 0.05$) either assuming a constant population size or after simulations considering the expansion process. A similar result was obtained when applying the different Fu and Li's tests (results not shown). However, Fu and Li's D and F tests using either *D. guanche* or *D. madeirensis* as outgroup were significant ($P < 0.05$) at *Psc* and *Ph-d* under the unrealistic assumption of a constant population size, which indicates the importance of using an accurate demographic model to calculate the significance of the neutrality tests.

Nucleotide variation in the coding region was also analyzed independently for synonymous (π_s) and nonsynonymous (π_a) sites (Table 2). *Caf1-55* with no nonsynonymous polymorphism and the lowest nonsynonymous divergence (K_a) is the most conserved PcG gene. Strong purifying selection acting on *Caf1-55* had been previously reported in a larger sample of *Drosophila* species (Calvo-Martín et al. 2016). On the other hand, a lack of synonymous polymorphism is present at *Pho*. In fact, all polymorphic sites in the *Pho* coding region present nonsynonymous variants. However, the overall nonsynonymous diversity at *Pho* is not high ($\pi_a = 0.0006$) given the low number of polymorphic sites (only 4). The highest nonsynonymous diversity is detected at *Psc* ($\pi_a = 0.0019$) that also present a high nonsynonymous/synonymous diversity ratio ($\pi_a/\pi_s = 0.1586$), which is only overcome by *Pcl* ($\pi_a/\pi_s = 0.1704$). *Pho*, *Pcl* and *Psc* also exhibit the three highest values of nonsynonymous divergence (K_a): 0.0181, 0.0129 and 0.0097, respectively.

Under strict neutrality, the ratio between intraspecific polymorphism and interspecific divergence is expected to be homogeneous across loci in a population of constant size at mutation-drift equilibrium. Based on this expectation, the HKA neutrality test (Hudson, Kreitman and Aguadé 1987) was developed to detect a putative decoupling in the polymorphism to divergence ratio between two gene regions. Assuming that a region evolves under neutrality, a significant excess of polymorphism at the other region would indicate balancing selection. Conversely, a recent selective sweep would result in a deficit of polymorphism. The multilocus HKA test performed considering silent variation in the sixteen PcG genes was not significant when *D. madeirensis* was used as outgroup ($\chi^2 = 10.86$, $df = 15$, $P = 0.3825$), suggesting homogeneity in the polymorphism to divergence ratio. However, the test was marginally significant when the outgroup was *D. guanche* ($\chi^2 = 22.09$, $df = 15$, $P = 0.0812$). This result stands out given that the HKA multilocus test is conservative when a single gene from a large sample present a ratio influenced by selection. A close inspection to the data considering divergence with *D. guanche* indicates that this is the case for the PcG genes studied since *Pho* deviates considerably from the

Table 1 Summary estimates of the level and pattern of polymorphism

Complex Gene	n° lines	n° sites	S	h	Hd	π	π/K	Tajima's D	R_m
PhoRC									
<i>Pho</i>	17	2842	11	10	0.90	0.0009	0.0263	-0.8763	0
<i>Sfmbt</i>	15	5617	76	14	0.99	0.0033	0.1055	-0.9055	9
<i>Phol</i>	16	2293	34	15	0.99	0.0030	0.1130	-1.3772	5
Pcl-PRC2									
<i>Caf1-55</i>	14	2621	85	13	0.99	0.0083	0.1857	-0.9447	9
<i>E(z)</i>	16	3652	143	15	0.99	0.0113	0.2162	-0.3766	16
<i>Esc</i>	15	2105	78	15	1.00	0.0098	0.3617	-0.7133	10
<i>Su(z)12</i>	16	3763	58	15	0.99	0.0039	0.1788	-0.7093	15
<i>Pcl</i>	16	3491	39	16	1.00	0.0027	0.0833	-0.8365	5
<i>Escl</i>	15	1756	66	14	0.99	0.0117	0.3400	-0.1027	10
PRC1									
<i>Psc</i>	15	11839	325	15	1.00	0.0060	0.1799	-1.3784	47
<i>Sce</i>	9	2714	50	9	1.00	0.0064	0.1455	-0.4042	6
<i>Pc</i>	17	3749	161	16	0.99	0.0085	0.2312	-1.4846	20
<i>Ph-p</i>	17	9203	176	16	0.99	0.0038	0.1348	-1.4433	20
<i>Scm</i>	16	4161	123	15	0.99	0.0060	0.1716	-1.4928	18
<i>Ph-d</i>	16	5752	63	15	0.99	0.0022	0.1007	-1.4728	6
dRAF*									
<i>Kdm2</i>	16	8507	186	15	0.99	0.0046	0.1403	-1.3780	20

S: number of segregating sites, *h*: number of haplotypes, *Hd*: haplotype diversity, π : nucleotide diversity, π/K : nucleotide diversity normalized by nucleotide divergence with *D. guanche*, R_m : minimum number of recombination events. *dRAF complex also contains PSC and SCE.

other genes. In fact, *Pho* has the largest contribution (54%) to the overall χ^2 -like test statistic and the contribution of the other PcG genes is quite uniform and small (Supplementary Figure S2). In addition, the high contribution of *Pho* is significant according to the maximum cell value test (Wang and Hey 1996). According to this result, the HKA test was applied in pairwise comparisons between *Pho* and each one of the other PcG genes. The 15 performed HKA tests were statistically significant ($P < 0.05$). This result clearly supports that *Pho* has a significant deficit of nucleotide polymorphism.

The McDonald and Kreitman (1991) test of neutrality (MK test) was also performed for each gene independently. This test is based on the comparison of the ratio of nonsynonymous to synonymous polymorphisms (P_a/P_s) and the ratio of nonsynonymous to synonymous fixed differences between species (F_a/F_s). Both ratios

Table 2 Summary of synonymous and nonsynonymous polymorphism and divergence with *D. guanche* in the PcG genes

Complex Gene	π_s	π_a	π_a / π_s	K_s	K_a	K_a / K_s
PhoRC						
<i>Pho</i>	0.0000	0.0006	-	0.0663	0.0181	0.2732
<i>Sfmbt</i>	0.0115	0.0009	0.0742	0.0890	0.0070	0.0786
<i>Phol</i>	0.0094	0.0012	0.1261	0.1051	0.0045	0.0428
Pcl-PRC2						
<i>Caf1-55</i>	0.0211	0.0000	0.0000	0.1417	0.0010	0.0070
<i>E(z)</i>	0.0242	0.0004	0.0145	0.1743	0.0036	0.0206
<i>Esc</i>	0.0234	0.0004	0.0176	0.0876	0.0022	0.0256
<i>Su(z)12</i>	0.0149	0.0006	0.0382	0.0621	0.0070	0.1133
<i>Pcl</i>	0.0071	0.0012	0.1704	0.0881	0.0129	0.1464
<i>Escl</i>	0.0334	0.0014	0.0407	0.1292	0.0034	0.0266
PRC1						
<i>Psc</i>	0.0120	0.0019	0.1586	0.0890	0.0097	0.1093
<i>Sce</i>	0.0149	0.0007	0.0443	0.0470	0.0033	0.0704
<i>Pc</i>	0.0128	0.0006	0.0494	0.0757	0.0061	0.0811
<i>Ph-p</i>	0.0070	0.0010	0.1355	0.0425	0.0070	0.1642
<i>Scm</i>	0.0137	0.0010	0.0736	0.0774	0.0063	0.0820
<i>Ph-d</i>	0.0056	0.0006	0.1097	0.0478	0.0092	0.1923
dRAF*						
<i>Kdm2</i>	0.0120	0.0006	0.0507	0.0901	0.0030	0.0334

π_s : synonymous nucleotide diversity, π_a : nonsynonymous nucleotide diversity, π_a / π_s : nonsynonymous/synonymous diversity ratio, K_s : synonymous divergence, K_a : nonsynonymous divergence, K_a / K_s : nonsynonymous/synonymous divergence ratio. *dRAF complex also contains PSC and SCE.

are expected to be equal under strict neutrality and a *G*-test of independence can be used to detect putative deviations from this expectation. A significant difference in both ratios was inferred only at *Su(z)12* when using *D. guanche* as outgroup ($G = 4.909$; 1 df, $P = 0.0267$). The direction of departure from neutrality can be inferred by the neutrality index (*NI*) estimated as the ratio between P_a/F_a and P_s/F_s (Rand and Kann 1996). Assuming that synonymous mutations are neutral, $NI > 1$ indicates an excess of nonsynonymous polymorphism and $NI < 1$ an excess of nonsynonymous divergence. The neutrality index at *Su(z)12* is 0.29, which indicates adaptive selection favouring nonsynonymous changes during the divergence of *D. subobscura* and *D. guanche*. The proportion of nonsynonymous substitutions fixed by natural selection at *Su(z)12* in these lineages is 0.71 according to the α -parameter (Smith and Eyre-Walker 2002).

DISCUSSION

Selection at the Polycomb group genes

Polycomb group (PcG) proteins form different complexes that maintain the repressive state of chromatin. Chromatin remodeling proteins are not generally considered proteins prone to have suffered adaptive changes. However, the action of adaptive selection has been inferred in different genes implicated in chromatin metabolism in *Drosophila*, as *rhino* that belongs to the heterochromatin protein 1 (HP1) family (Vermaak, Henikoff and Malik 2005). More surprising, genome wide analysis performed to detect gene regions with a strong genetic differentiation between tropical and temperate populations of *D. melanogaster* also identified chromatin organization genes such as *crm* (Gibert, Karch and Schlötterer 2011), *Ph-p* (Voigt et al. 2015), *chd1*, *ssrp*, *chm* and *glu* (Levine and Begun 2008).

Adaptive selection and coevolution has also been inferred at different PcG proteins during the divergence of the *Drosophila* genus (Calvo-Martín et al. 2016), especially in the branch ancestral to the species of the *obscura* group. In contrast to this previous study, the present analysis of nucleotide variation in *D. subobscura* in sixteen PcG genes failed to detect signals of positive selection in most of them. In fact, the footprint of adaptive selection could be only detected in *Su(z)12* during the divergence of *D. subobscura* and *D. guanche*. This gene also showed an acceleration of the nonsynonymous substitution rate in the branch ancestral to the *obscura* group species (Calvo-Martín et al. 2016). SU(Z)12 is a noncatalytic subunit of the PRC2 Polycomb complex that potentiates the enzymatic activity of its partner E(Z) and contributes to nucleosome binding (Ketel et al. 2005; Nekrasov, Wild and Müller 2005). The protein has three domains: the N-terminal domain that interacts with CAF1-55, a zinc finger C2H2 and the VEFS box. Most of the fourteen amino acid replacements detected in the SU(Z)12 protein of *D. subobscura* and *D. guanche* are located at the poorly conserved C-terminal region of the protein and none of them affect the described protein domains. Therefore, the functional relevance of these amino acid replacements is uncertain.

Nucleotide variation at the dot-like chromosome of *D. subobscura*

The Polycomb gene *Pho* is located at the dot-like chromosome in *D. subobscura* near the distal or telomeric end of the chromosome. This position is shared by *D. melanogaster*, *D. yakuba*, *D. teissieri* and *D. erecta*, which supports that this is the ancestral location in the *Sophophora* subgenus. Therefore, the proximal

location of *Pho* detected in other species of the *melanogaster* group such as *D. orena*, *D. mauritiana*, *D. simulans* and *D. sechellia* would be the result of chromosomal inversions as suggested by Podemski, Ferrer and Locke (2001).

Nucleotide variation at *Pho* in *D. subobscura* when compared to other Polycomb genes is peculiar in several aspects: low nucleotide diversity, low number of haplotypes, rather low haplotype diversity and only nonsynonymous polymorphism. In addition, no recombination events are detected by the four gamete test and thus linkage disequilibrium estimated by $|D'|$ is equal to 1 in all pairwise comparisons between parsimony informative sites. Therefore, variation at *Pho* would support that the dot-like chromosome of *D. subobscura* also has an exceedingly low recombination rate as previously reported in other *Drosophila* species such as *D. melanogaster* (Sandler and Szauter 1978). This lack of recombination has a drastic effect on nucleotide variation. In fact, several multilocus studies have confirmed that the level and pattern of nucleotide variation differ between the dot-like chromosome and the other autosomes in *D. melanogaster* (Wang et al. 2002) and other *Drosophila* species not only of the *melanogaster* group such as *D. simulans* (Wang et al. 2004) and *D. yakuba* (Arguello et al. 2010) but also of the *virilis* group as *D. americana* (Betancourt, Welch and Charlesworth 2009). These studies show a low level of nucleotide polymorphism and relatively inefficient selection in the genes located in the dot-like chromosome.

The low level of variation in this chromosome is explained by both the hitchhiking effect (Maynard Smith and Haigh 1974) and the background selection model (Charlesworth, Morgan and Charlesworth 1993). Thus, in regions with low recombination, the fixation of positively selected mutations and the elimination of deleterious mutations cause a reduction of neutral linked variants. This reduction of variation can also be explained by the Hill-Robertson effect (Hill and Robertson 1966) that proposes that a locus linked to a second locus under directional selection experiences a reduction in effective population size (N_e). The decrease in N_e , apart from reducing neutral variation, causes a relaxation of selection due to an increase in the effects of genetic drift. The presence of only nonsynonymous polymorphism at the *Pho* gene of *D. subobscura* is consistent with a relaxation of selection against slightly deleterious nonsynonymous mutations. In addition, the low codon bias at *Pho* would indicate inefficient selection acting at synonymous sites to maintain the prevalent use of particular synonymous codons (Akashi 1995).

In summary, the level and patterns of nucleotide variation at the PcG genes in *D. subobscura* support the action of positive selection only at *Su(z)12*. The comparison of the ratio between polymorphic and fixed sites differs significantly for synonymous and nonsynonymous variants, suggesting adaptive evolution during the divergence of *D. subobscura* and *D. guanche*. In addition, the characteristics of the nucleotide polymorphism at *Pho* are consistent with the location of this gene in the dot-like chromosome and would indicate that this chromosome also exhibits little or no recombination in *D. subobscura*.

ACKNOWLEDGEMENTS

We thank Serveis Científic-Tècnics, Universitat de Barcelona, for automated DNA sequencing facilities. This work was supported by a predoctoral fellowship from Agència de Gestió d'Ajuts Universitaris i de Recerca (AGAUR), Generalitat de Catalunya, Catalonia, Spain, to JM C-M.; and grants BFU2012-35168 and BFU2015-63732 from Ministerio de Economía y Competitividad, Spain, and 2009SGR-1287 and 2014SGR10555 from Comissió Interdepartamental de Recerca i Innovació Tecnològica, Generalitat de Catalunya, Catalonia, Spain, to Montserrat Aguadé.

REFERENCES

- Akashi H (1995) Inferring weak selection from patterns of polymorphism and divergence at silent sites in *Drosophila* DNA. *Genetics* 139: 1067–1076.
- Alvarez-Ponce D, Guirao-Rico S, Orengo DJ, Segarra C, Rozas J et al. (2012) Molecular population genetics of the insulin/TOR signal transduction pathway: a network-level analysis in *Drosophila melanogaster*. *Mol. Biol. Evol.* 29: 123–132.
- Arguello JR, Zhang Y, Kado T, Fan C, Zhao R et al. (2010) Recombination yet inefficient selection along the *Drosophila melanogaster* subgroup's fourth chromosome. *Mol. Biol. Evol.* 27: 848–861.
- Beck EA, Llopart A (2015) Widespread positive selection drives differentiation of centromeric proteins in the *Drosophila melanogaster* subgroup. *Sci. Rep.* 5: 17197.
- Begun DJ, Holloway AK, Stevens K, Hillier LW, Poh Y-P et al. (2007) Population genomics: whole-genome analysis of polymorphism and divergence in *Drosophila simulans*. *PLoS Biol.* 5: e310.
- Betancourt AJ, Welch JJ, Charlesworth B (2009) Reduced effectiveness of selection caused by a lack of recombination. *Curr. Biol.* 19: 655–660.

- Burland TG (2000) DNASTAR's Lasergene sequence analysis software. *Methods Mol. Biol.* 132: 71–91.
- Calvo-Martín JM, Librado P, Aguadé M, Papaceit M, Segarra C (2016) Adaptive selection and coevolution at the proteins of the Polycomb repressive complexes in *Drosophila*. *Heredity* 116: 213–223.
- Calvo-Martín JM, Papaceit M, Segarra C (2017) Evidence of neofunctionalization after the duplication of the highly conserved Polycomb group gene *Caf1-55* in the obscura group of *Drosophila*. *Sci. Rep.* 7: 40536.
- Charlesworth B, Morgan MT, Charlesworth D (1993) The effect of deleterious mutations on neutral molecular variation. *Genetics* 134: 1289–1303.
- Clark NL, Aquadro CF (2010) A novel method to detect proteins evolving at correlated rates: identifying new functional relationships between coevolving proteins. *Mol. Biol. Evol.* 27: 1152–1161.
- Cork JM, Purugganan MD (2004) The evolution of molecular genetic pathways and networks. *BioEssays* 26: 479–484.
- Eanes WF (2011) Molecular population genetics and selection in the glycolytic pathway. *J. Exp. Biol.* 214: 165–171.
- Findlay GD, Sitnik JL, Wang W, Aquadro CF, Clark NL et al. (2014) Evolutionary rate covariation identifies new members of a protein network required for *Drosophila melanogaster* female post-mating responses. *PLoS Genet.* 10: e1004108.
- Flores HA, DuMont VLB, Fatoo A, Hubbard D, Hijji M et al. (2015) Adaptive evolution of genes involved in the regulation of germline stem cells in *Drosophila melanogaster* and *D. simulans*. *G3* 5: 583–592.
- Fu YX, Li WH (1993) Statistical tests of neutrality of mutations. *Genetics* 133: 693–709.
- Gibert JM, Karch F, Schlötterer C (2011) Segregating variation in the Polycomb group gene *cramped* alters the effect of temperature on multiple traits. *PLoS Genet.* 7: e1001280.
- Hill WG, Robertson A (1966) The effect of linkage on limits to artificial selection. *Genet. Res.* 8: 269–294.
- Hudson RR, Kaplan NL (1985) Statistical properties of the number of recombination events in the history of a sample of DNA sequences. *Genetics* 111: 147–164.
- Hudson RR, Kreitman M, Aguadé M (1987) A test of neutral molecular evolution based on nucleotide data. *Genetics* 116: 153–159.
- Ketel CS, Andersen EF, Vargas ML, Strome S, Simon JA (2005) Subunit contributions to histone methyltransferase activities of fly and worm Polycomb group complexes. *Mol. Cell. Biol.* 25: 6857–6868.
- Langley CH, Stevens K, Cardeno C, Lee YCG, Schrider DR et al. (2012) Genomic variation in natural populations of *Drosophila melanogaster*. *Genetics* 192: 533–598.
- Lanzuolo C, Orlando V (2012) Memories from the Polycomb group proteins. *Annu. Rev. Genet.* 46: 561–589.

- Leung W, Shaffer CD, Cordonnier T, Wong J, Itano MS et al. (2010) Evolution of a distinct genomic domain in *Drosophila*: comparative analysis of the dot chromosome in *Drosophila melanogaster* and *Drosophila virilis*. *Genetics* 185: 1519–1534.
- Levine MT, Begun DJ (2008) Evidence of spatially varying selection acting on four chromatin-remodeling loci in *Drosophila melanogaster*. *Genetics* 179: 475–485.
- Lewontin RC (1964) The interaction of selection and linkage. I. General considerations; heterotic models. *Genetics* 49: 49–67.
- Librado P, Rozas J (2009) DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics* 25: 1451–1452.
- Mackay TFC, Richards S, Stone EA, Barbadilla A, Ayroles JF et al. (2012) The *Drosophila melanogaster* genetic reference panel. *Nature* 482: 173–178.
- Maynard Smith J, Haigh J (1974) The hitch-hiking effect of a favourable gene. *Genet. Res.* 23: 23–35.
- McDonald JH, Kreitman M (1991) Adaptive protein evolution at the *Adh* locus in *Drosophila*. *Nature* 351: 652–654.
- Nekrasov M, Wild B, Müller J (2005) Nucleosome binding and histone methyltransferase activity of *Drosophila* PRC2. *EMBO Rep.* 6: 348–353.
- Pál C, Papp B, Lercher MJ (2006) An integrated view of protein evolution. *Nat. Rev. Genet.* 7: 337–348.
- Papaceit M, Segarra C, Aguadé M (2013) Structure and population genetics of the breakpoints of a polymorphic inversion in *Drosophila subobscura*. *Evolution* 67: 66–79.
- Podemski L, Ferrer C, Locke J (2001) Whole arm inversions of chromosome 4 in *Drosophila* species. *Chromosoma* 110: 305–312.
- Pratdesaba R, Segarra C, Aguadé M (2015) Inferring the demographic history of *Drosophila subobscura* from nucleotide variation at regions not affected by chromosomal inversions. *Mol. Ecol.* 24: 1729–1741.
- Presgraves DC, Stephan W (2007) Pervasive adaptive evolution among interactors of the *Drosophila* hybrid inviability gene, *Nup96*. *Mol. Biol. Evol.* 24: 306–314.
- Ramos-Onsins SE, Mitchell-Olds T (2007) Mlcoalsim: multilocus coalescent simulations. *Evol. Bioinform. Online* 3: 41–44.
- Rand DM, Kann LM (1996) Excess amino acid polymorphism in mitochondrial DNA: contrasts among genes from *Drosophila*, mice, and humans. *Mol. Biol. Evol.* 13: 735–748.
- Sackton TB, Lazzaro BP, Schlenke TA, Evans JD, Hultmark D et al. (2007) Dynamic evolution of the innate immune system in *Drosophila*. *Nat. Genet.* 39: 1461–1468.
- Sandler L, Szauter P (1978) The effect of recombination-defective meiotic mutants on fourth-chromosome crossing over in *Drosophila melanogaster*. *Genetics* 90: 699–712.
- Smith NGC, Eyre-Walker A (2002) Adaptive protein evolution in *Drosophila*. *Nature* 415: 1022–1024.
- Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 123: 585–595.

- Tamura K, Stecher G, Peterson D, Filipski A, Kumar S (2013) MEGA6: Molecular evolutionary genetics analysis version 6.0. *Mol. Biol. Evol.* 30: 2725–2729.
- Vermaak D, Henikoff S, Malik HS (2005) Positive selection drives the evolution of *rhino*, a member of the heterochromatin protein 1 family in *Drosophila*. *PLoS Genet.* 1: 96–108.
- Vitti JJ, Grossman SR, Sabeti PC (2013) Detecting natural selection in genomic data. *Annu. Rev. Genet.* 47: 97–120.
- Voigt S, Laurent S, Litovchenko M, Stephan W (2015) Positive selection at the *polyhomeotic* locus led to decreased thermosensitivity of gene expression in temperate *Drosophila melanogaster*. *Genetics* 200: 591–599.
- Wang RL, Hey J (1996) The speciation history of *Drosophila pseudoobscura* and close relatives: inferences from DNA sequence variation at the period locus. *Genetics* 144: 1113–1126.
- Wang W, Thornton K, Berry A, Long M (2002) Nucleotide variation along the *Drosophila melanogaster* fourth chromosome. *Science* 295: 134–137.
- Wang W, Thornton K, Emerson JJ, Long M (2004) Nucleotide variation and recombination along the fourth chromosome in *Drosophila simulans*. *Genetics* 166: 1783–1794.
- Wong A, Turchin MC, Wolfner MF, Aquadro CF (2008) Evidence for positive selection on *Drosophila melanogaster* seminal fluid protease homologs. *Mol. Biol. Evol.* 25: 497–506.
- Wright F (1990) The ‘effective number of codons’ used in a gene. *Gene* 87: 23–29.

Molecular population genetics of the Polycomb genes in
Drosophila subobscura

Calvo-Martín JM, Papaceit M, Segarra C

Supplementary Material

Table S1 Location of the PcG genes on the chromosomes of *D. subobscura* (A, J, U, E, O and dot-like) and chromosomal arrangements of the sequenced lines. The dot-like chromosome has not chromosomal polymorphism.

Complex Gene	Chromosome	Chromosomal section	Arrangements
PhoRC			
Pho	Dot-like	100 D	-
Sfmbt	U	41 C	U ₁₊₂
Phol	J	22 D	J ₁
Pcl-PRC2			
Caf1-55	O	79 C	O _{st} , O ₃₊₄ , O ₃₊₄₊₈
E(z)	J	21 C	J _{st} , J ₁
Esc	U	49 A	U ₁₊₂
Su(z)12	J	25 C	J _{st} , J ₁
Pcl	E	56 B	E _{st} , E ₁₊₂ , E ₁₊₂₊₉₊₁₂
Escl	U	47 B	U ₁₊₂
PRC1			
Psc	E	66 C	E _{st} , E ₁₊₂
Sce	O	96 A	O _{st}
Pc	J	33 A	J _{st} , J ₁
Ph-p	A	13 A	A _{st} , A ₁
Scm	O	77 D	O _{st} , O ₃₊₄ , O ₃₊₄₊₈
Ph-d	A	13 A	A _{st} , A ₁
dRAF			
Kdm2	O	82 B	O _{st} , O ₃₊₄ , O ₃₊₄₊₈

Figures S1a to S1p

Nucleotide polymorphic sites of each of the sixteen PcG genes studied in the *D. subobscura* sequenced lines. The blue bar above the alignment illustrates gene structure: 5' flanking region (5'), coding exons numbered by order (E), introns also numbered by order (I) and 3' flanking region (3'). The identification of each line is given on the left by the letters OF (Observatori Fabra) and a number. Sites are numbered according to the multiple alignment. Asterisks under site numbers indicate nonsynonymous polymorphic sites. Dots indicate nucleotides identical to the first sequence that is used as reference. Gaps of a single nucleotide are shown by dashes. Insertions and deletions are indicated by the letters *i* and *d*, respectively, with a number according to its length. Polymorphic microsatellites are shown in brackets followed by a number that indicates the number of repeats. The last two rows of the multiple alignment show the information for the polymorphic sites in the *D. madeirensis* and *D. guanche* sequences.

Esc

5'										E1	I1	I2	E3	I3	E4	3'			
OF01	C	A	T	T	C	Cl ₄	T	C	C	T	C	G	T	C	A	T	A	A	-
OF07	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
OF13	-	T	C	-	Cl ₃	-	T	A	-	-	-	-	-	-	-	-	-	-	-
OF14	-	T	C	-	Cl ₃	-	C	A	C	T	-	G	T	C	-	-	-	-	-
OF15	T	T	C	-	Cl ₃	-	A	T	-	G	A	-	-	-	-	-	-	-	-
OF16	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
OF18	-	-	-	-	Cl ₄	-	-	-	-	-	-	-	-	-	-	-	-	-	-
OF19	-	T	C	A	Cl ₁	C	T	C	-	A	-	-	-	-	-	-	-	-	-
OF21	T	T	C	-	Cl ₁	C	T	C	-	C	A	T	C	G	-	-	-	-	-
OF23	-	d ₃	C	-	Cl ₄	-	-	-	-	-	-	-	-	-	-	-	-	-	-
OF31	T	T	C	-	Cl ₁	C	T	C	-	A	T	-	-	-	-	-	-	-	-
OF42	-	T	C	A	Cl ₁	C	T	C	-	A	T	-	-	-	-	-	-	-	-
OF47	-	T	C	A	Cl ₁	C	T	C	-	A	T	-	-	-	-	-	-	-	-
OF49	-	-	-	-	Cl ₄	-	-	-	-	-	-	-	-	-	-	-	-	-	-
OF60	-	T	C	-	Cl ₁	C	T	C	-	A	T	-	-	-	-	-	-	-	-
OF74	-	T	C	-	Cl ₃	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Dmad	-	T	C	T	Cl ₃	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Dgua	-	T	-	-	-	Cl ₃	-	-	-	-	-	-	-	-	-	-	-	-	-

Figure S1c

Esci/

	5'	E1	I1	E2	I2	E3	3'
OF01	T T G A	(A) ₁	T A G T T C A G C G G G A T C G C C A A T C T G C G G C C A G C G C C C C T C T C G G T C G G T T A T G A C C G C				1759
OF07	. C C .	(A) ₁	. T .	. C A G C A T T .	. G G C .	. T T G A .	1758
OF13	. C C .	(A) ₁	. T .	. C A G C A T T .	. G G C .	. T T G A .	1753
OF14	. . C T (A) ₁	. . G .	. . A .	. . C A G C A T T .	. . G G C .	. . T T G A .	1751
OF15	. . C T (A) ₁	. . G .	. . A .	. . C A G C A T T .	. . G G C .	. . T T G A .	1712
OF18	. . C T (A) ₁	. . G .	. . A .	. . C A G C A T T .	. . G G C .	. . T T G A .	1706
OF19	. . C T (A) ₁	. . G .	. . A .	. . C A G C A T T .	. . G G C .	. . T T G A .	1692
OF21	. . C T (A) ₁	. . G .	. . A .	. . C A G C A T T .	. . G G C .	. . T T G A .	1681
OF23	. . C T (A) ₁	. . G .	. . A .	. . C A G C A T T .	. . G G C .	. . T T G A .	1669
OF31	. . C T (A) ₁	. . G .	. . A .	. . C A G C A T T .	. . G G C .	. . T T G A .	1668
OF42	. . C T (A) ₁	. . G .	. . A .	. . C A G C A T T .	. . G G C .	. . T T G A .	1663
OF47	. . C T (A) ₁	. . G .	. . A .	. . C A G C A T T .	. . G G C .	. . T T G A .	1486
OF49	. . C T (A) ₁	. . G .	. . A .	. . C A G C A T T .	. . G G C .	. . T T G A .	1482
OF60	. . C T (A) ₁	. . G .	. . A .	. . C A G C A T T .	. . G G C .	. . T T G A .	1441
OF74	. . C T (A) ₁	. . G .	. . A .	. . C A G C A T T .	. . G G C .	. . T T G A .	1402
Dmad	. . C .	(A) ₁	. . T .	. . C A G .	. . T .	. . T .	1378
Dgua	. . A .	Ag(A) ₂	. . T .	. . T .	. . T .	. . T .	1252

Figure S1d

PCl

	E1	E2	E3	E4	3'
100	AATG	TTGG	CCGC	GC	ATG
99	G	C	C	A	G
98	G	C	C	A	G
97	G	C	C	A	G
96	G	C	C	A	G
95	G	C	C	A	G
94	G	C	C	A	G
93	G	C	C	A	G
92	G	C	C	A	G
91	G	C	C	A	G
90	G	C	C	A	G
89	G	C	C	A	G
88	G	C	C	A	G
87	G	C	C	A	G
86	G	C	C	A	G
85	G	C	C	A	G
84	G	C	C	A	G
83	G	C	C	A	G
82	G	C	C	A	G
81	G	C	C	A	G
80	G	C	C	A	G
79	G	C	C	A	G
78	G	C	C	A	G
77	G	C	C	A	G
76	G	C	C	A	G
75	G	C	C	A	G
74	G	C	C	A	G
73	G	C	C	A	G
72	G	C	C	A	G
71	G	C	C	A	G
70	G	C	C	A	G
69	G	C	C	A	G
68	G	C	C	A	G
67	G	C	C	A	G
66	G	C	C	A	G
65	G	C	C	A	G
64	G	C	C	A	G
63	G	C	C	A	G
62	G	C	C	A	G
61	G	C	C	A	G
60	G	C	C	A	G
59	G	C	C	A	G
58	G	C	C	A	G
57	G	C	C	A	G
56	G	C	C	A	G
55	G	C	C	A	G
54	G	C	C	A	G
53	G	C	C	A	G
52	G	C	C	A	G
51	G	C	C	A	G
50	G	C	C	A	G
49	G	C	C	A	G
48	G	C	C	A	G
47	G	C	C	A	G
46	G	C	C	A	G
45	G	C	C	A	G
44	G	C	C	A	G
43	G	C	C	A	G
42	G	C	C	A	G
41	G	C	C	A	G
40	G	C	C	A	G
39	G	C	C	A	G
38	G	C	C	A	G
37	G	C	C	A	G
36	G	C	C	A	G
35	G	C	C	A	G
34	G	C	C	A	G
33	G	C	C	A	G
32	G	C	C	A	G
31	G	C	C	A	G
30	G	C	C	A	G
29	G	C	C	A	G
28	G	C	C	A	G
27	G	C	C	A	G
26	G	C	C	A	G
25	G	C	C	A	G
24	G	C	C	A	G
23	G	C	C	A	G
22	G	C	C	A	G
21	G	C	C	A	G
20	G	C	C	A	G
19	G	C	C	A	G
18	G	C	C	A	G
17	G	C	C	A	G
16	G	C	C	A	G
15	G	C	C	A	G
14	G	C	C	A	G
13	G	C	C	A	G
12	G	C	C	A	G
11	G	C	C	A	G
10	G	C	C	A	G

Figure S1f

Pho

	5'	I1	E2	I4	E5								
2692						0969							
2690						0080							
2581						2581							
187						565							
1509						1509							
1348						1348							
1140						1140							
1064						1064							
838						838							
53						53							
366						366							
300						300							
252						252							
122						122							
Chou	A	T	-	C	G	A	A	G	G	G	A	A	*
OF01	.	.	.	.	C	G	.	.	.	.	.	.	.
OF07	.	.	.	.	C	G	.	.	.	.	.	.	.
OF14	.	.	.	.	C	G	.	.	.	.	.	.	.
OF18	.	.	.	.	C	.	.	.	.	.	.	.	.
OF19	.	.	.	.	.	.	A	.	.	.	.	.	.
OF21	.	.	.	.	C	.	.	.	.	.	.	.	.
OF23	G	C	.	C	.	.	.	T	.	.	.	.	.
OF28	G	.	.	C	.	.	.	T	.	.	.	.	.
OF31	.	.	.	C	.	.	.	.	.	.	.	.	.
OF40	G	.	.	C	.	.	.	T	.	.	.	.	.
OF47	G	.	.	d ₁	C	.	.	T	.	.	.	.	.
OF49	.	.	.	C	.	.	.	T	.	.	.	.	.
OF54	.	.	.	C	.	.	.	.	.	.	.	.	.
OF58	.	.	.	C	.	.	.	.	.	.	.	.	.
OF60	.	.	.	.	.	.	A	.	.	.	.	.	.
OF78	.	d ₄	.	k	.	.	C	.	.	.	.	.	.
Dmad	?	?	?	?	.	C	.	.	.	.	.	.	.
Dgua	?	?	?	?	.	C	.	.	.	.	.	.	.

Figure S1g

Phol

	5'	E1	
OF07	CGCCG	CATCC	CGTCA
OF14	CGCCG	CATCC	CGTCA
OF15	CGCCG	CATCC	CGTCA
OF19	CGCCG	CATCC	CGTCA
OF23	CGCCG	CATCC	CGTCA
OF26	CGCCG	CATCC	CGTCA
OF28	CGCCG	CATCC	CGTCA
OF31	CGCCG	CATCC	CGTCA
OF40	CGCCG	CATCC	CGTCA
OF42	CGCCG	CATCC	CGTCA
OF44	CGCCG	CATCC	CGTCA
OF58	CGCCG	CATCC	CGTCA
OF60	CGCCG	CATCC	CGTCA
OF74	CGCCG	CATCC	CGTCA
OF77	CGCCG	CATCC	CGTCA
OF78	CGCCG	CATCC	CGTCA
Dmad	CGCCG	CATCC	CGTCA
Dgua	CGCCG	CATCC	CGTCA

Figure S1h

Sfmbt

	5'	E1	I1	E2	I2	E3	E4	I4	E5	I5	E6	I6	E7	3'			
OF01	T	G	A	G	G	A	G	G	C	A	G	A	G	G	T	A	(AC) ₁₀
OF07	T	T	A	G	C	A	G	C	A	G	A	G	A	G	A	A	(AC) ₁₀
OF13	T	T	A	G	C	A	G	C	A	G	A	G	A	G	A	A	(AC) ₁₀
OF14	T	T	A	G	C	A	G	C	A	G	A	G	A	G	A	A	(AC) ₁₀
OF15	T	T	A	G	C	A	G	C	A	G	A	G	A	G	A	A	(AC) ₁₀
OF18	T	T	A	G	C	A	G	C	A	G	A	G	A	G	A	A	(AC) ₁₀
OF19	T	T	A	G	C	A	G	C	A	G	A	G	A	G	A	A	(AC) ₁₀
OF21	T	T	A	G	C	A	G	C	A	G	A	G	A	G	A	A	(AC) ₁₀
OF23	G	T	A	G	C	A	G	C	A	G	A	G	A	G	A	A	(AC) ₁₀
OF31	T	T	A	G	C	A	G	C	A	G	A	G	A	G	A	A	(AC) ₁₀
OF42	T	T	A	G	C	A	G	C	A	G	A	G	A	G	A	A	(AC) ₁₀
OF47	T	T	A	G	C	A	G	C	A	G	A	G	A	G	A	A	(AC) ₁₀
OF49	T	T	A	G	C	A	G	C	A	G	A	G	A	G	A	A	(AC) ₁₀
OF60	T	T	A	G	C	A	G	C	A	G	A	G	A	G	A	A	(AC) ₁₀
OF74	T	T	A	G	C	A	G	C	A	G	A	G	A	G	A	A	(AC) ₁₀
Dmad	T	T	A	G	C	A	G	C	A	G	A	G	A	G	A	A	(AC) ₁₀
Dgua	T	T	A	G	C	A	G	C	A	G	A	G	A	G	A	A	(AC) ₁₀

Figure S1i

Psc

		5'		E1		I1	
Chou	OF01	18	121	A G C G C A A d ₁ C C G C C C G A	- A A C T T G C A T A T C G G T A G A A d ₁ h ₈ G C C T T C G A A A T C A h ₉ G A C	- G A - A T G T C G G C A C T T C	T C A G G G
OF01	OF01			- A - C -	- G - G - T -	- C A -	- T G - A A
OF07	OF07			- A - C -	- G - G - T -	- C A -	- T G - A A
OF14	OF14			- A - C -	- G - G - T -	- C A -	- T G - A A
OF15	OF15			- A - C -	- G - G - T -	- C A -	- T G - A A
OF18	OF18			- A - C -	- G - G - T -	- C A -	- T G - A A
OF21	OF21			- A - C -	- G - G - T -	- C A -	- T G - A A
OF23	OF23			- A - C -	- G - G - T -	- C A -	- T G - A A
OF28	OF28			- A - C -	- G - G - T -	- C A -	- T G - A A
OF31	OF31			- A - C -	- G - G - T -	- C A -	- T G - A A
OF34	OF34			- A - C -	- G - G - T -	- C A -	- T G - A A
OF47	OF47			- A - C -	- G - G - T -	- C A -	- T G - A A
OF49	OF49			- A - C -	- G - G - T -	- C A -	- T G - A A
OF74	OF74			- A - C -	- G - G - T -	- C A -	- T G - A A
OF78	OF78			- A - C -	- G - G - T -	- C A -	- T G - A A
Dmad	Dmad			- A - C -	- G - G - T -	- C A -	- T G - A A
Dgua	Dgua			- A - C -	- G - G - T -	- C A -	- T G - A A

		E2		I2	
Chou	OF01	2309	2645	4159	4341
OF01	OF01			4159	4341
OF07	OF07			4159	4341
OF14	OF14			4159	4341
OF15	OF15			4159	4341
OF18	OF18			4159	4341
OF21	OF21			4159	4341
OF23	OF23			4159	4341
OF28	OF28			4159	4341
OF31	OF31			4159	4341
OF34	OF34			4159	4341
OF47	OF47			4159	4341
OF49	OF49			4159	4341
OF74	OF74			4159	4341
OF78	OF78			4159	4341
Dmad	Dmad			4159	4341
Dgua	Dgua			4159	4341

Figure S1j

continued on next page

continued from previous page

Psc

E4		3'	
10994	* G C C C C G G A G A C A T	12190	G A C C G C T C A A T A A T
11020	* A T	12191	C T
11032	* T	12200	C T
11046	*	12207	C T
11191	*	12209	C T
11251	*	12341	C T
11290	*	12372	C T
11476	*	12373	C T
11568	*	12393	C T
11581	*	12425	C T
11804	*	12432	C T
11854	*	12498	C T
11879	*	12533	C T
12013	*	12551	C T
12078	*	12594	C T
12121	*	12627	C T
Chou			
OF01			
OF07			
OF14			
OF15			
OF18			
OF21			
OF23			
OF28			
OF31			
OF40			
OF47			
OF49			
OF74			
OF78			
Dmad			
Dgua			

Figure S1j

Pc

	5'										3'									
Chou	T	T	T	G	G	A	G	A	A	T	T	A	A	A	G	G	T	A	T	T
OF07	G	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
OF14	G	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
OF15	G	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
OF19	G	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
OF23	G	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
OF26	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
OF28	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
OF31	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
OF33	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
OF40	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
OF42	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
OF44	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
OF58	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
OF80	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
OF74	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
OF77	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
OF78	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
Dmad	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
Dgua	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.

	5'										3'									
Chou	A	C	G	C	A	A	C	A	T	G	A	A	A	A	G	T	T	A	T	T
OF07	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
OF14	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
OF15	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
OF19	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
OF23	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
OF26	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
OF28	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
OF31	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
OF40	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
OF42	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
OF44	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
OF58	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
OF80	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
OF74	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
OF77	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
OF78	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
Dmad	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
Dgua	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.

Figure S11

continued from previous page

Ph-p

		E4		I4	E5		3'	
7438								
7444								
7521								
7537								
7599								
7626								
7659								
7758								
7910								
8026								
8029								
8047								
8059								
8081								
8093								
8146								
8151								
8190								
8225								
8227								
8241								
8274								
8283								
8328								
8342								
8380								
8580								
8678								
8743								
8860								
8862								
8876								
9092								
9101								
9107								
9134								
9217								
9236								
9280								
9305								
9317								
9632								
9665								
9677								
9746								
9750								
9774								
9808								
9837								
9842								
9855								
9960								
9961								
9981								
10000								
10010								
Chou								
OF07								
OF14								
OF15								
OF16								
OF18								
OF19								
OF23								
OF28								
OF40								
OF42								
OF44								
OF47								
OF49								
OF58								
OF60								
OF74								
Dmad								
Dgua								

Figure S1m

Scm

	5'	E1	E2	I1	I2	I3	E4	3'
5	113	113	113	1020	1042	1049	1052	1053
61	113	113	113	1020	1042	1049	1052	1053
62	113	113	113	1020	1042	1049	1052	1053
63	113	113	113	1020	1042	1049	1052	1053
64	113	113	113	1020	1042	1049	1052	1053
65	113	113	113	1020	1042	1049	1052	1053
66	113	113	113	1020	1042	1049	1052	1053
67	113	113	113	1020	1042	1049	1052	1053
68	113	113	113	1020	1042	1049	1052	1053
69	113	113	113	1020	1042	1049	1052	1053
70	113	113	113	1020	1042	1049	1052	1053
71	113	113	113	1020	1042	1049	1052	1053
72	113	113	113	1020	1042	1049	1052	1053
73	113	113	113	1020	1042	1049	1052	1053
74	113	113	113	1020	1042	1049	1052	1053
75	113	113	113	1020	1042	1049	1052	1053
76	113	113	113	1020	1042	1049	1052	1053
77	113	113	113	1020	1042	1049	1052	1053
78	113	113	113	1020	1042	1049	1052	1053
79	113	113	113	1020	1042	1049	1052	1053
80	113	113	113	1020	1042	1049	1052	1053
81	113	113	113	1020	1042	1049	1052	1053
82	113	113	113	1020	1042	1049	1052	1053
83	113	113	113	1020	1042	1049	1052	1053
84	113	113	113	1020	1042	1049	1052	1053
85	113	113	113	1020	1042	1049	1052	1053
86	113	113	113	1020	1042	1049	1052	1053
87	113	113	113	1020	1042	1049	1052	1053
88	113	113	113	1020	1042	1049	1052	1053
89	113	113	113	1020	1042	1049	1052	1053
90	113	113	113	1020	1042	1049	1052	1053
91	113	113	113	1020	1042	1049	1052	1053
92	113	113	113	1020	1042	1049	1052	1053
93	113	113	113	1020	1042	1049	1052	1053
94	113	113	113	1020	1042	1049	1052	1053
95	113	113	113	1020	1042	1049	1052	1053
96	113	113	113	1020	1042	1049	1052	1053
97	113	113	113	1020	1042	1049	1052	1053
98	113	113	113	1020	1042	1049	1052	1053
99	113	113	113	1020	1042	1049	1052	1053
100	113	113	113	1020	1042	1049	1052	1053
101	113	113	113	1020	1042	1049	1052	1053
102	113	113	113	1020	1042	1049	1052	1053
103	113	113	113	1020	1042	1049	1052	1053
104	113	113	113	1020	1042	1049	1052	1053
105	113	113	113	1020	1042	1049	1052	1053
106	113	113	113	1020	1042	1049	1052	1053
107	113	113	113	1020	1042	1049	1052	1053
108	113	113	113	1020	1042	1049	1052	1053

Figure S10

continued from previous page

Kdm2

	3'
8188	G A A T C A A T C - A T T T G G C G C - T A C T T C T T
8189	
8195	
8206	
8215	
8220	
8224	
8243	
8296	
8338	
8382	
8390	
8436	
8441	
8442	
8457	
8462	
8464	
8468	
8494	
8495	
8525	
8534	
8542	
8568	
8572	
8687	
8700	
Chcu	
OF01	
OF07	
OF14	
OF16	
OF18	
OF19	
OF21	
OF28	
OF31	
OF40	
OF47	
OF54	
OF58	
OF60	
OF78	
d ₁	
d ₂	
Dmad	
Dgua	

Figure S1p

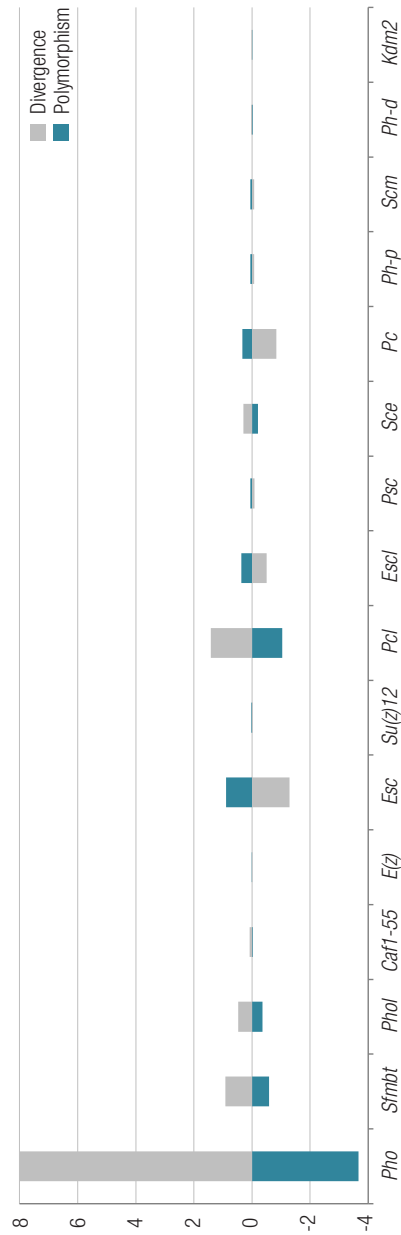


Figure S2. Contribution of each PcG gene to the multilocus HKA test. Gray bars indicate the contribution to the overall χ^2 -like test statistic due to silent divergence between *D. subobscura* and *D. guanche*. Blue bars indicate the corresponding contribution due to silent polymorphism. The highest contribution of *Pho* both to divergence and polymorphism is significant by the maximum cell value test (see text).

Artículo 4

Sex combs reduced (Scr) regulatory region of *Drosophila* revisited

Calvo-Martín JM, Papaceit M, Segarra C

2017. *Molecular Genetics and Genomics*. doi: 10.1007/s00438-017-1309-1

El gen Hox *Sex combs reduced (Scr)* es responsable de la diferenciación de los segmentos labial y protorácico en *Drosophila*. *Scr* se expresa en varios tejidos específicos durante el desarrollo embrionario, siguiendo un complejo programa que debe ser coordinado, a su vez, por una región reguladora igualmente compleja. Aunque ya se han descrito algunos módulos reguladores en *cis* (CRMs) en la región *upstream* de *Scr* (~75kb), no se ha realizado un estudio detallado y sistemático de los distintos elementos reguladores que la componen. En este trabajo, la región reguladora de *Scr* ha sido revisada con el objetivo de solventar esta carencia. El estudio se ha centrado en la identificación de *Initiator elements* (IEs), a los que se unen los factores de segmentación, y *Polycomb response elements* (PREs), reconocidos por los complejos Polycomb y Trithorax, así como *insulators* y *tethering elements*. Para ello, se ha recopilado toda la información disponible en la actualidad sobre la región reguladora de *Scr*, obtenida principalmente en proyectos a nivel genómico de ChIP-chip y ChIP-seq. Además, se ha realizado un análisis bioinformático de la región en busca de IEs y PREs mediante el programa MOTEVO, basado en la conservación evolutiva de las secuencias reguladoras. Los resultados obtenidos con esta estrategia combinada son ampliamente consistentes con los CRMs identificados en trabajos previos y han contribuido a: i) delimitarlos con más precisión, ii) subdividir dos de ellos en elementos independientes, iii) identificar uno nuevo, iv) determinar los sitios de unión que los componen y v) redefinir algunas de sus características. Los buenos resultados obtenidos indican que aproximaciones como la utilizada en este trabajo, integrando datos funcionales y análisis bioinformático, pueden ser de gran utilidad a la hora de caracterizar otras regiones reguladoras.

Sex combs reduced (*Scr*) regulatory region of *Drosophila* revisited

Juan M. Calvo-Martín¹ · Montserrat Papaceit¹ · Carmen Segarra¹ 

Received: 5 September 2016 / Accepted: 8 March 2017
© Springer-Verlag Berlin Heidelberg 2017

Abstract The Hox gene *Sex combs reduced* (*Scr*) is responsible for the differentiation of the labial and prothoracic segments in *Drosophila*. *Scr* is expressed in several specific tissues throughout embryonic development, following a complex path that must be coordinated by an equally complex regulatory region. Although some cis-regulatory modules (CRMs) have been identified in the *Scr* regulatory region (~75 kb), there has been no detailed and systematic study of the distinct regulatory elements present within this region. In this study, the *Scr* regulatory region was revisited with the aim of filling this gap. We focused on the identification of Initiator elements (IEs) that bind segmentation factors, Polycomb response elements (PREs) that are recognized by the Polycomb and Trithorax complexes, as well as insulators and tethering elements. To this end, we summarized all currently available information, mainly obtained from high throughput ChIP data projects. In addition, a bioinformatic analysis based on the evolutionary conservation of regulatory sequences using the software MOTEVO was performed to identify IE and PRE candidates in the *Scr* region. The results obtained by this combined strategy are largely consistent with the CRMs previously identified in the *Scr* region and help to: (i) delimit them more accurately,

(ii) subdivide two of them into different independent elements, (iii) identify a new CRM, (iv) identify the composition of their binding sites and (v) better define some of their characteristics. These positive results indicate that an approach that integrates functional and bioinformatic data might be useful to characterize other regulatory regions.

Keywords *Sex combs reduced* (*Scr*) · Cis-regulatory modules · Initiator elements · Polycomb response elements · *Drosophila*

Introduction

The Homeotic selector genes (Hox) encode a set of transcription factors essential in the process of embryonic development. These genes determine the identity and subsequent development of the different segments along the anterior-posterior axis of the embryo (Hughes and Kaufman 2002). This capacity to decide cell fate makes them key players in morphological evolution, not only in arthropods, in which they have been mostly studied, but also in all bilaterians (Pick and Heffer 2012).

The proper temporal and spatial expression of the Hox genes depends on a cascade of transcription factors encoded by the so-called segmentation genes, which generate a progressive and accurate expression pattern along the anteroposterior axis of the embryo. This cascade begins with the maternal genes already present in the oocyte cytoplasm and continues with the gap genes, the pair-rule genes, and the segment polarity genes (Peel et al. 2005). These segmentation factors direct the expression of Hox genes in their respective action domains by binding to cis-regulatory modules (CRMs) called Initiator elements (IEs) (Maeda and Karch 2006).

Communicated by S. Hohmann.

Electronic supplementary material The online version of this article (doi:10.1007/s00438-017-1309-1) contains supplementary material, which is available to authorized users.

✉ Carmen Segarra
csegarra@ub.edu

¹ Departament de Genètica, Microbiologia i Estadística, Facultat de Biologia, i Institut de Recerca de la Biodiversitat (IRBio), Universitat de Barcelona, Barcelona, Spain

Segmentation genes, however, only dictate where each Hox gene is expressed or repressed initially. Two new classes of regulators with antagonistic functions are in charge of maintaining the correct transcriptional state during the rest of embryonic development and adult life: the Polycomb group genes (PcG) that maintain transcriptionally silent states (Lanzuolo and Orlando 2012) and the Trithorax group genes (TrxG) that maintain transcriptionally active states (Schuettengruber et al. 2011; Steffen and Ringrose 2014). Polycomb and Trithorax group proteins form complexes that undertake post-translational modifications of histones, thereby changing the conformation of chromatin and altering both the accessibility of the transcriptional machinery (Müller and Verrijzer 2009) and the three-dimensional structure of the regulated loci, which can group together and form the so-called Polycomb bodies (Bantignies and Cavalli 2011).

Polycomb and Trithorax complexes recognize target genes by means of CRMs called Polycomb response elements (PREs) (Kassis and Brown 2013). The only members of the Polycomb and Trithorax groups with DNA binding activity and, therefore, able to bind to PREs are the Polycomb proteins Pleiohomeotic (PHO) (Brown et al. 1998; Fritsch et al. 1999; Wang et al. 2004) and Pleiohomeotic like (PHOL) (Brown et al. 2003), which are encoded by paralogous genes. However, a set of additional cofactors is required to ensure the proper recruitment of the repressive and activating complexes. The best characterized among these cofactors are the following: Z (Hur et al. 2002), TRL (a.k.a. Gaga factor) (Mahmoudi et al. 2003), DSP1 (Déjardin et al. 2005), PSQ (Americo et al. 2002), SPSS (Brown and Kassis 2010), GRH (Blastyák et al. 2006), and ADF1 (Orsi et al. 2014).

Neither IEs nor PREs have a common sequence that allows their identification by simple homology. Each CRM is unique and its ability to recruit transcription factors or the PcG and TrxG complexes appears to differ according to the requirements of each regulated gene (Okulski et al. 2011). In fact, analysis of divergence among these elements in several *Drosophila* species showed that the evolutionary constraints do not act at the sequence level, but by maintaining the necessary number of transcription factor binding sites (TFBSs) and their proper spatial distribution, whether these regulatory motifs are homologous between species or have arisen de novo (Hauenschild et al. 2008; Starr et al. 2012).

In recent years, in silico methods to identify CRMs regardless of sequence similarity have been developed. These methods are mainly based on a sliding window process to detect clusters of a predefined set of TFBSs (Berman et al. 2002; Ringrose et al. 2003). However, they have been further refined to take into account the conservation relative to other species (Berman et al. 2004;

Schroeder et al. 2004), to improve the weight matrix of each TFBS by training sets (Fiedler and Rehmsmeier 2006), and lastly to develop a full probabilistic framework that takes into account the genomic or evolutionary context (Arnold et al. 2012; Zeng et al. 2012).

Another approach to detect CRMs is the genome-wide direct mapping of regulatory proteins using chromatin immunoprecipitation techniques (ChIP). Those DNA regions in which several proteins involved in a particular regulatory system colocalize are clear CRM candidates. In *Drosophila*, the modENCODE project (Celniker et al. 2009; Roy et al. 2010; Nègre et al. 2011) and other independent studies (MacArthur et al. 2009; Schuettengruber et al. 2009) have provided a large amount of information in this regard.

One of the *Drosophila* Hox genes known to be under Polycomb/Trithorax regulation is *Sex combs reduced* (*Scr*), which is responsible for the differentiation of the labial and prothoracic segments in *Drosophila*. The *Scr* gene, which is part of the Antennapedia (*Antp*) complex (Kaufman et al. 1980), is composed of four exons (Kuroiwa et al. 1985). The first two noncoding exons are incorporated alternatively in the two transcripts generated by the two different transcription start sites (Lemotte et al. 1989; Andrew 1995). The *Scr* regulatory region extends ~75 kb and encompasses the pair-rule gene *Fushi tarazu* (*Ftz*). *Scr* is expressed in several specific tissues throughout embryonic development and follows a complex pathway. In the early stages of the embryogenesis, the *Scr* transcript appears at the epidermal level in the labial and prothoracic segments (parasegments 2 and 3). Subsequently, it can be found at the mesoderm level in the prothoracic segment, as well as in the gastric caeca of the midgut, the labial neuromere of the central nervous system and a narrow band of cells in the ventral region of the subesophageal ganglion. Finally, *Scr* is expressed in the labial and prothoracic imaginal discs of pupae and in the corresponding organs of the adult flies (Mahaffey and Kaufman 1987; Riley et al. 1987; Kaufman et al. 1990; Reuter and Scott 1990).

The complex expression pattern that is characteristic of Hox genes must be supported by an equally complex regulatory region. This region was studied by T. Kaufman's group two decades ago in *Drosophila melanogaster* (Gindhart and Kaufman 1995; Gindhart et al. 1995; Gorman and Kaufman 1995) with the identification of eight candidate regulatory elements (*Scr* CRMs). However, this region has not been examined in depth since then, with the exception of the regulatory fragment corresponding to *Ftz* (Calhoun and Levine 2003), and the identification of a couple of tethering elements (Calhoun et al. 2002; Calhoun and Levine 2003) and two insulators (Belozeroev et al. 2003; Li et al. 2015).

In this study, we revisited the *D. melanogaster* *Scr* regulatory region with the aim of collecting and integrating all currently available data that have been obtained both by specific studies and high throughput projects, mainly those based on ChIP data. In addition, we reanalyzed this regulatory region using the CRM detection software MOTEVO, which enabled new insights into the evolutionary conservation of each of the identified regulatory elements. To this end, we used as inputs the alignment of the *Scr* regulatory region in eight *Drosophila* species and a complete set of TFBSs of the proteins involved in the regulation of *Scr* by binding to IEs or PREs. The results obtained by this combined approach are consistent with most of the CRMs previously identified and help better define the characteristics of many of them. Therefore, the approach performed herein might prove useful for detailed study of other regulatory regions.

Materials and methods

High throughput data

The ChIP-chip data for the segmentation transcription factors used to identify IEs were obtained from the Berkeley *Drosophila* Transcription Network Project (BDTNP) (Li et al. 2008; MacArthur et al. 2009) (<http://bdtnp.lbl.gov/Fly-Net>). The ChIP-chip and ChIP-seq data for the regulatory proteins associated with PREs and other elements, as insulators, were retrieved from the modENCODE *D. melanogaster* GBrowse (Roy et al. 2010) (<http://www.modencode.org>) and from Schuettengruber et al. (2009). In addition, we compiled data on the expression patterns of candidate enhancers in *D. melanogaster* obtained in the Fly Enhancers Project (Kvon et al. 2014) (<http://enhancers.starklab.org>).

In order to detect IE candidates from ChIP data, we focused on the binding profiles of the products of the segmentation genes *Bicoid* (*Bcd*), *Caudal* (*Cad*), *Giant* (*Gt*), *Hunchback* (*Hb*), *Knirps* (*Kni*), *Kruppel* (*Kr*), *Tailless* (*Tll*), *Dichaete* (*D*), *Huckebein* (*Hkb*), *Fushi tarazu* (*Ftz*), *Hairless* (*H*), *Paired* (*Prd*), *Runt* (*Run*), and *Sloppy paired 1* (*Slp1*). In order to infer PREs, the binding profiles studied were those of the PcG proteins encoded by *Pleiohomeotic* (*Pho*), *Polyhomeotic* (*Ph*), *Scm-related gene containing four MBT domains* (*Sfmbt*), *Enhancer of zeste* (*E(z)*), *Polycomb-like* (*Pcl*), *Polycomb* (*Pc*), *Posterior sex combs* (*Psc*), and *Sex combs extra* (*Sce*), as well as those of the TrxG proteins encoded by *Trithorax* (*Trx*), *Brahma* (*Brm*), *Nejire* (*Nej*), *Enhancer of bithorax* (*E(bx)*), *Imitation switch* (*Iswi*), *Mi-2*, *Absent small or homeotic discs 1* (*Ash1*), and *Snf5-related 1* (*Snr1*). Finally, we also analyzed the binding profiles of the three cofactors encoded by *Trithorax-like*

(*Trl*), *Zeste* (*Z*), and *Dorsal switch protein 1* (*Dsp1*) as well as the profiles of five histone deacetylases (HDAC1, 3, 4, 6, and 11) and five proteins involved in the formation of insulators, which are encoded by *Modifier of mdg4* (*Mod(mdg4)*), *Suppressor of Hairy wing* (*Su(Hw)*), *Centrosomal protein 190kD* (*Cp190*), *CTCF*, and *Boundary element-associated factor of 32kD* (*BEAF-32*).

Bioinformatic prediction of CRMs

Sequences of the *Scr* regulatory region in eight *Drosophila* species (*D. melanogaster*, *D. simulans*, *D. sechellia*, *D. yakuba*, *D. erecta*, *D. ananassae*, *D. pseudoobscura*, and *D. persimilis*) were retrieved from FlyBase release 6.09 (<http://flybase.org>) and multiple-aligned using the MUSCLE program (Edgar 2004) implemented in the software MEGA 6 (Tamura et al. 2013). MEGA 6 was also used to infer by maximum likelihood the branch lengths of the phylogenetic tree reconstructed from divergence in this region according to the well-established phylogenetic relationships of the studied species.

The resulting alignment and the reconstructed phylogenetic tree were used as inputs in the software MOTEVO (Arnold et al. 2012), which was developed to infer the presence of CRMs in a multiple alignment using a Bayesian probabilistic framework. In addition, a set of TFBS weight matrixes for each protein involved in the regulation of *Scr* was used as a third input. These weight matrixes were obtained from the Fly Factor Survey (FFS) database (<http://pgfe.umassmed.edu/ffs>), and those absent from FFS were inferred from the bibliography.

In order to infer IEs by MOTEVO, the TFBSs for the same set of segmentation factors analyzed in the high throughput ChIP data analysis were used. Regarding PREs, the set of TFBSs includes, besides the binding sites for the already mentioned PHO, TRL, Z, and DSP1, those for the proteins involved in the PRE-mediated regulation encoded by the genes *Pipsqueak* (*Psq*), *Sp1-like factor for pairing sensitive silencing* (*Spps*), *Grainy head* (*Grh*), *Adh transcription factor 1* (*Adf1*), and binding site A (bsA) (Brown and Kassiss 2013).

MOTEVO was also used to identify the binding sites for each transcription factor in a 1 kb region around each CRM candidate and to calculate the posterior probability for the identified TFBS. Moreover, the binding capacity for each transcription factor in this 1 kb region relative to the whole *Scr* regulatory region was inferred. With this purpose and by means of a sliding window process (1 kb windows with a step size of 1 bp), a distribution for the sum of the posterior probabilities of the binding sites for each transcription factor present in each 1 kb window across the *Scr* regulatory region was constructed. Then the percentile rank of the

window including a CRM predicted by MOTEVO was estimated from this distribution.

Results

High throughput ChIP data

ChIP-chip and ChIP-seq data were collected to infer the location of two classes of CRMs: the Initiator elements (IEs) involved in the initial expression of the Hox genes in the appropriate segment during the first stages of the embryonic development, and the Polycomb/Trithorax response elements (PREs) responsible for the maintenance of the Hox genes' transcriptional state during subsequent stages of development and in adult flies. In order to infer IEs, the binding profiles of the proteins encoded by two early maternal, seven gap, and five pair-rule genes were analyzed. In order to infer PREs, we focused on the binding profiles of eight Polycomb and eight Trithorax proteins, as well as those of three known PRE cofactors. In addition, we also studied the binding profiles of five histone deacetylases and five proteins associated with insulator elements (see “Material and methods” section for details).

Figure 1 shows the ChIP data collected for the *Scr* regulatory region. Those positions at which a group of proteins clearly colocalize are framed and labeled from A to M. The most noteworthy pattern is the clear difference between frames where the segmentation factors (early maternal, gap, and pair-rule) colocalize and frames enriched by the PcG- and TrxG-related proteins. The only exception is frame A where the segmentation factors colocalize with PHO, PH, low levels of several other PcG and TrxG proteins, the cofactors TRL, Z, and DSP1 and histone deacetylases 1, 4, and 6. Among the frames occupied solely by segmentation factors, and therefore candidates to constitute IEs, there are also differences relative to the proportion of gap and pair-rule proteins that bind to them. In fact, frames E and I are composed mainly by gap proteins, while frame F and especially H are composed mainly by pair-rule proteins. Both gap and pair-rule proteins are present in frames G and L. With respect to the early maternal proteins, CAD is present in all frames occupied by the other segmentation factors, while BCD seems to be absent from frames A and L and appears at low levels in frames E and F.

Frames B, C, and M predict regulatory regions that bind not only to PcG and TrxG proteins, but also to their cofactors TRL, Z, and DSP1. Therefore, they are good candidates for functional PREs. These three frames are also targeted by the histone deacetylases 1, 4, and 6, and the insulator-related protein MOD (MDG4). There are no remarkable differences between these three frames beyond the apparent absence of TRX in frames B and C or the lack of DSP1

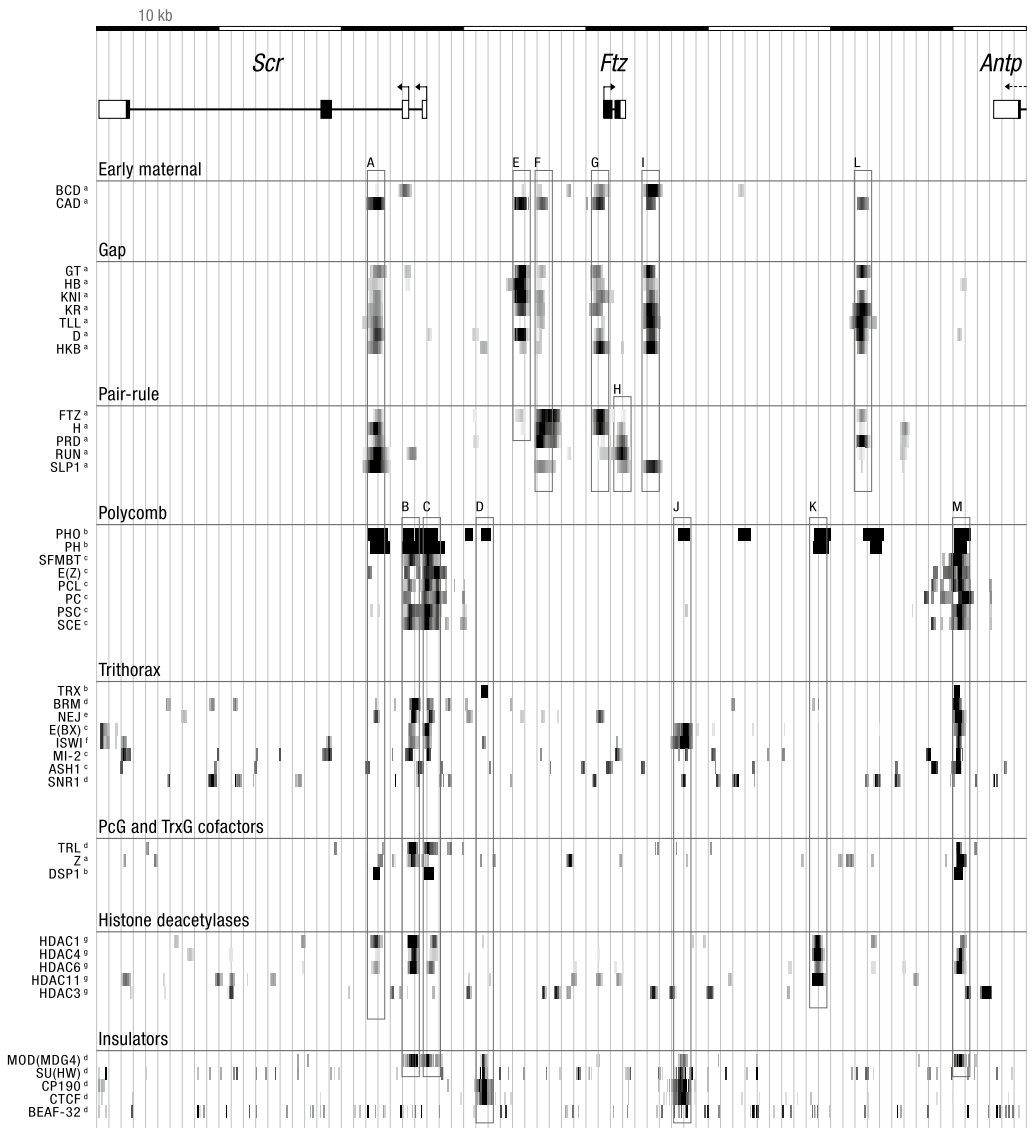
Fig. 1 Summary of the available ChIP data for the *Scr* regulatory region in *D. melanogaster*. The diagrams at the top show gene location, organization, and size along the ~75 kb studied region according to the displayed scale. Coding exons are indicated by black boxes, noncoding exons by white boxes, and introns by thin lines. Arrows pinpoint the transcription start site and direction. The binding profiles for each regulatory protein along the *Scr* region are represented using a gray scale from white to black according to the binding scores. White areas correspond to binding scores lower than the 90th percentile rank with respect to all scores in the analyzed region, and black areas to scores greater than the 99th percentile rank. DNA fragments that stand out for having high binding scores for different sets of transcription factors are framed and henceforth named frames A to M. The source data for each transcription factor are as follows: (a) Berkeley *Drosophila* Transcription Network Project, ChIP-chip data from 1 h 20'–2 h 50' embryos; (b) Schuettengruber group, ChIP-chip data from 4 to 12 h embryos (as these data only indicate the presence/absence of significant binding, no gray scale is shown); (c–h) *Drosophila* modENCODE project; (c) Karpen group, ChIP-chip data from S2-DRSC cells; (d) White group, ChIP-chip data from 0 to 12 h embryos; (e) White group, ChIP-chip data corresponding to the mean from 0 to 4, 4–8, 8–12, 12–16, 16–20, and 20–24 h embryos; (f) Karpen group, ChIP-chip data from Kc167 cells; (g) White group, ChIP-seq data from 0 to 12 h embryos. See text for references

in frame B. On the other hand, MOD (MDG4) also colocalizes in frames D and J with the rest of the components of the insulator complexes, except BEAF-32, which has a diffuse binding profile. Therefore, frames D and J are insulator candidates, although these frames also colocalize the PcG protein PHO and several TrxG proteins, mainly TRX in frame D, and E(BX) and ISWI in frame J. Finally, frame K is the only region enriched in histone deacetylases that includes HDAC11 and is depleted of PcG (except PHO and PH) and TrxG proteins, thus probably being a regulatory element distinct from IEs or PREs.

Frames A to M were identified by high throughput ChIP data obtained in *D. melanogaster*. These kind of functional data are not available in other *Drosophila* species and thus it cannot be analyzed whether these frames are evolutionary conserved. However, different bioinformatic approaches have been developed to predict CRMs and one of these approaches (MOTEVO) relies on the phylogenetic conservation of the target DNA sequences known to bind specific transcription factors. Thus, MOTEVO was applied to further corroborate the results of high throughput ChIP data analysis in the *Scr* regulatory region in *D. melanogaster* and to gain new insights into the conservation of the *Scr* CRMs in other species.

Bioinformatic detection of CRMs

Starting with the multiple alignment of the *Scr* regulatory region in eight *Drosophila* species of the Sophophora subgenus, MOTEVO was used to predict CRMs by means of two different sets of TFBs. First, to infer IEs, we used a set of TFBs corresponding to the same 14 segmentation



factors studied in the ChIP analysis. Second, to infer PREs, we used the set of TFBSs corresponding to the Polycomb DNA binding protein PHO, the three cofactors TRL, Z, and DSP1, as well as five PRE-related proteins.

Figure 2 illustrates the enhancer scores provided by MOTEVO along the entire *Scr* regulatory region of *D. melanogaster* independently for IE and PRE candidates. The ten major peaks for each CRM class are labeled according

to their scores and named henceforth IE1 to IE10 and PRE1 to PRE10. The exact location of these CRM candidates in the *D. melanogaster* genome is provided in Online Resource 1. In the case of IE candidates, peaks IE1 and IE2 are clearly higher than the other peaks and overlap perfectly with ChIP frames I and E (Fig. 1), respectively, which are occupied by segmentation factors. Peaks IE3 and IE4 have their counterpart in frames L and A, respectively, which

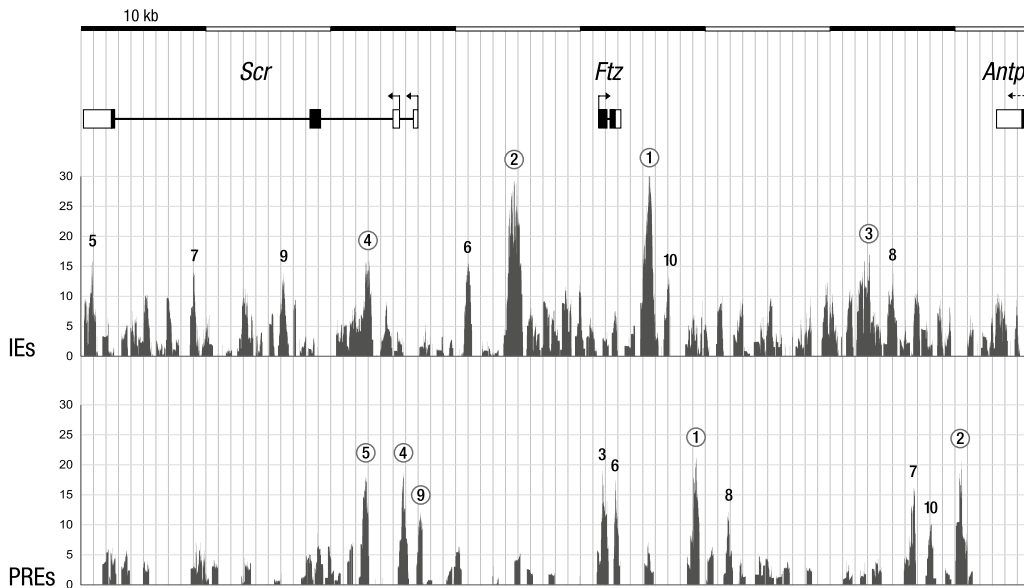


Fig. 2 Bioinformatic detection of CRMs along the *Scr* regulatory region of *D. melanogaster* with MOTEVO. The gene content of the studied region is shown in the top part of the figure (see Fig. 1). Each graph presents the enhancer scores (y-axis) provided by MOTEVO along the *Scr* regulatory region (x-axis). The first graph shows the predicted IEs using a set of TFBSs for segmentation factors and the

graph below the predicted PREs using a set of TFBSs for proteins involved in the PcG/TrxG-mediated regulation. The ten major peaks inferred from each analysis are labeled from 1 to 10 according to scores and those that are the best candidates for relevant functional CRMs (see text) are circled

are also enriched in segmentation factors. The remaining peaks, IE5–IE10, have similar scores but do not colocalize with any ChIP frame. However, it is worth noting that the Berkeley *Drosophila* Transcription Network ChIP database was obtained from early embryos at stages 4 and 5 (from 1 h 20' to 2 h 50'). Although segmentation factors mainly direct embryonic development at early stages, they are also involved in several patterning processes at later stages (Isshiki et al. 2001). Therefore, it remains possible that peaks IE5 to IE10 are actually functional CRMs, even though they do not overlap with any ChIP frame.

Regarding PRE candidates, peaks PRE2 and PRE4 colocalize with ChIP frames M and B (Fig. 1), which are highly enriched in PcG and TrxG proteins, supporting the notion that they are functional PREs. Peak PRE9 colocalizes with ChIP frame C, which is also highly enriched in PcG and TrxG proteins. This fact and its location, like PRE4, just before one of the two transcription start sites of *Scr*, makes PRE9 a strong candidate for a functional PRE despite the rather low enhancer score given by MOTEVO. Peak PRE5 partially overlaps with peak IE4 and, therefore, colocalizes with ChIP frame A, which is occupied by both segmentation factors and a few PcG- and TrxG-related

proteins. In contrast, peak PRE1 does not colocalize with any ChIP frame, but its high score might not be consistent with it being a false positive. The nearby peaks PRE3 and PRE6 fall into the *Ftz* gene region and colocalize with ChIP frames G and H, which are associated with segmentation factors but not PcG and TrxG proteins, as would be expected for a PRE. Therefore, PRE3 and PRE6 are likely functional CRMs associated with *Ftz* but not *Scr* and so they will not be considered further. Finally, peaks PRE7, PRE8, and PRE10 do not colocalize with any ChIP frame.

MOTEVO provides phylogenetic support for the CRMs detected as it uses a multiple alignment as input. Therefore, the main IE and PRE candidates inferred in the *D. melanogaster* sequence are expected to be conserved in the other *Drosophila* species included in the multiple alignment. To further confirm this prediction, the MOTEVO analysis was repeated seven times, each time using a different *Drosophila* species as the reference. The results show that all the main peaks detected in *D. melanogaster* (Fig. 2) are also inferred in the other *Drosophila* species, with the only exception of peak PRE9, which seems to be absent in *D. ananassae*, *D. pseudoobscura*, and *D. persimilis* (Online Resource 2). Therefore, this analysis indicates that the main

Scr regulatory elements have been conserved during the divergence of *D. melanogaster* and *D. pseudoobscura* (or *D. persimilis*) that began ~24 million years ago (Obbard et al. 2012).

TFBS composition of the CRMs

MOTEVO was also used to identify the binding sites of each transcription factor and their posterior probabilities in a 1 kb window around the detected CRMs. This and subsequent analyses focused on the most relevant CRMs identified (labeled in a circle in Fig. 2), which are those predicted by MOTEVO that overlap with ChIP data frames. PRE1 was also included given its high score. Figure 3 shows the TFBSs detected for these IE and PRE candidates. In addition, this figure indicates the binding capacity of a

transcription factor in a CRM reflected by the percentile rank of the sum of the TFBS posterior probabilities in this CRM relative to the whole *Scr* regulatory region.

The TFBS composition differs among CRMs despite some shared characteristics. The two IEs with the highest enhancer scores, IE1 and IE2 (Fig. 2), are characterized by a pronounced enrichment and binding capacity for BCD and D, but they also present clear differences: IE1 is distinguished by its high occupancy of KR and TLL binding sites and IE2 by the predominance of CAD, HB, and KNI binding capacity. IE3 stands out by presenting the highest percentile rank of FTZ, and it is also enriched in CAD and D binding sites. IE4 shows the highest binding capacity for GT and RUN, and, like IE2, is also enriched in CAD and D binding sites.

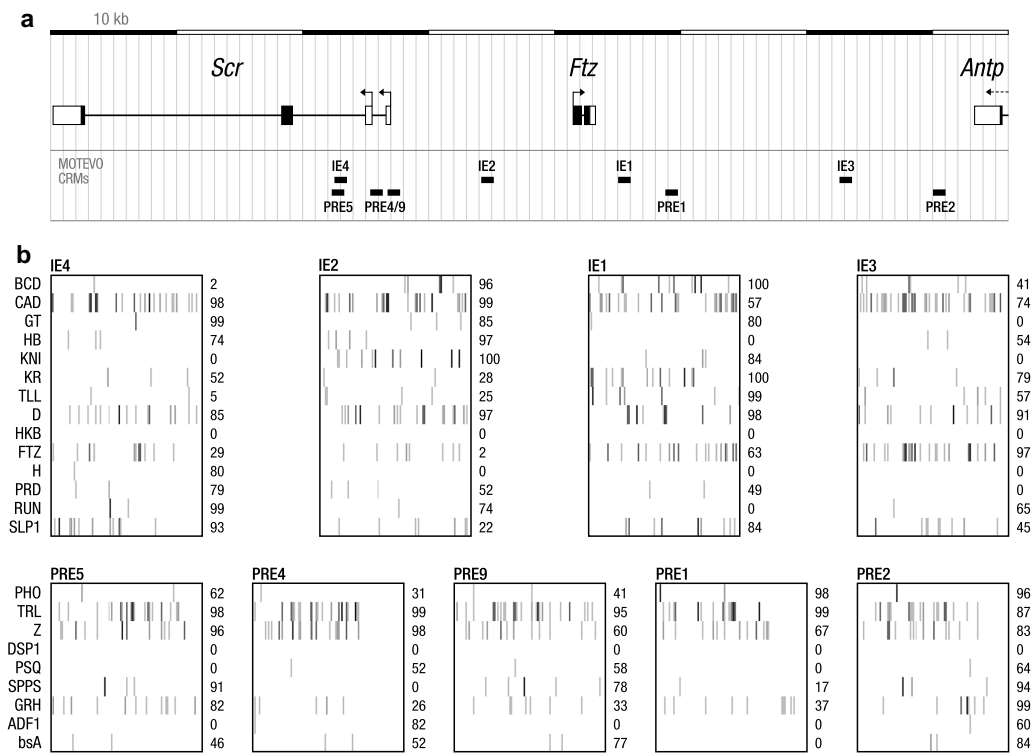


Fig. 3 Location and binding site composition of the relevant CRM candidates predicted by MOTEVO. **a** Below the diagram of the genes (see Fig. 1), the location of the main CRMs inferred by the MOTEVO bioinformatic analysis is indicated by thick black lines. Each CRM extends 1 kb around the site with the maximum enhancer score in the signal peak. **b** Along each CRM, the binding sites for the different transcription factors are indicated by vertical lines. The intensity of

these lines follows a gray scale according to the posterior probability of the TFBS: the darker the line, the higher the posterior probability. The numbers on the right of each CRM indicate the percentile rank of the sum of the binding sites' posterior probabilities for a transcription factor in that CRM, relative to the whole *Scr* regulatory region (see "Materials and methods" section)

The identified PREs also show similarities beside differences in their TFBS composition. At least one PHO binding site is present in each PRE candidate, which is expected given that PHO is indispensable for proper PRE-mediated regulation. The five PRE candidates are also rich in binding sites with a high percentile rank for the cofactors TRL and Z, while the binding capacity for the other proteins that mediate PRE regulation is more variable. PRE2 is the most enriched in TFBSs, showing a high binding capacity for SPPS, GRH, and bsA. On the contrary, PRE1 stands out as presenting a low number of SPPS and GRH binding sites and no binding sites for DSP1, PSQ, ADF1, or bsA. Despite the absence of binding sites for these transcription factors, PRE1 has the highest enhancer score in the MOTEVO analysis, which can be mainly attributed to the high percentile ranks for PHO and TRL.

Identification of tethering elements

MOTEVO was also used to find out whether the TTCGAA repeats that characterize the two tethering elements, TE and DTE (Distal TE), previously identified in *D. melanogaster* (Calhoun et al. 2002; Calhoun and Levine 2003), are present in the other *Drosophila* species studied herein. The results confirm that the homologous fragments of the TEs in the *Scr* regulatory region of these species are also enriched in motifs that perfectly match TTCGAA or differ from it by a single degenerate position. It is noteworthy that these TEs colocalize perfectly with PRE2 and PRE4 and, therefore, they are not independent regulatory elements but part of these PREs. Figure 4 shows the TTCGAA motifs identified in the 1 kb window assigned for PRE2 (DTE) and PRE4 (TE). The eight perfect TTCGAA repeats present in *D. melanogaster* PRE4 are conserved in *D. simulans* and *D. sechellia*, but reduced to seven in *D. yakuba*, *D. erecta*, and *D. ananassae* and to five in *D. pseudoobscura* and *D. persimilis*. However, in the last two species a new perfect repeat is present in the 1 kb window. In the case of PRE2, the five perfect repeats present in *D. melanogaster* are conserved in *D. simulans* and *D. sechellia*, reduced to four in *D. yakuba*, *D. erecta*, and *D. ananassae*, to two in *D. pseudoobscura*, and to just one in *D. persimilis*. However, one new TTCGAA repeat is present in *D. ananassae*, two in *D. pseudoobscura*, and three in *D. persimilis*. Therefore, PRE2 and PRE4 have at least four TTCGAA repeats in all species, this being the minimum number necessary for proper tethering function (Calhoun and Levine 2003).

In order to establish whether the presence of tethering elements is a common characteristic of PREs, MOTEVO was used to identify clusters of at least four TTCGAA motifs in 1 kb windows along the Antennapedia (Ant-C) and Bithorax (Bx-C) complexes of Hox genes in *D. melanogaster*. The Ant-C region extends over ~400 kb and

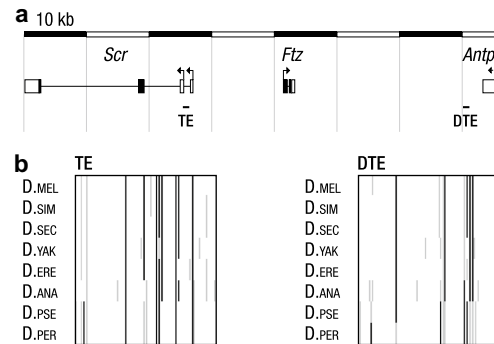


Fig. 4 Location of the tethering elements in the *Scr* regulatory region and motif conservation among species. **a** The two tethering elements, TE and DTE, are indicated by thick black lines below the diagram of the genes (see Fig. 1). As TE and DTE are included in PRE4 and PRE2, respectively, they are represented by the same 1 kb window as these PRE candidates. **b** The presence of the TTCGAA motif along this 1 kb window is indicated for each studied *Drosophila* species. Black lines correspond to exact TTCGAA motifs, and gray lines to one-nucleotide degenerated motifs. D.MEL, *D. melanogaster*; D.SIM, *D. simulans*; D.SEC, *D. sechellia*; D.YAK, *D. yakuba*; D.ERE, *D. erecta*; D.ANA, *D. ananassae*; D.PSE, *D. pseudoobscura*, and D.PER, *D. persimilis*

encompasses *Labial* (*Lab*), *Proboscipedia* (*Pb*), *Deformed* (*Df*), *Scr*, and *Antp*. The Bx-C region extends over ~400 kb and includes *Ultrabithorax* (*Ubx*), *Abdominal A* (*abd-A*), and *Abdominal B* (*Abd-B*). In addition to the *Scr* TE (PRE4) and DTE (PRE2), one cluster with four motifs in Ant-C and four clusters also with four motifs in Bx-C were identified. However, none of these clusters colocalize with known PREs or regions enriched in Polycomb proteins according to the modENCODE ChIP data.

Expression profiles

As a last step to further characterize the CRMs present in the *Scr* regulatory region, the expression profiles of putative enhancers reported by Kvon et al. (2014) and available at <http://enhancers.starklab.org> were analyzed. This database provides information on the expression during embryogenesis of a Gal4 reporter under the control of a ~2 kb noncoding DNA fragment randomly chosen from the genome of *D. melanogaster*. Figure 5 shows the studied enhancer candidates present in the *Scr* regulatory region and the expression profiles of those that colocalize with the most relevant CRMs identified herein. As expected, the two enhancer candidates that include IEs drive expression mainly in the early embryo (stage 4–6). Thus, VT37565 and VT37571 that overlap IE2 and IE1, respectively, induce expression in two anteroposterior stripes resembling a subset of

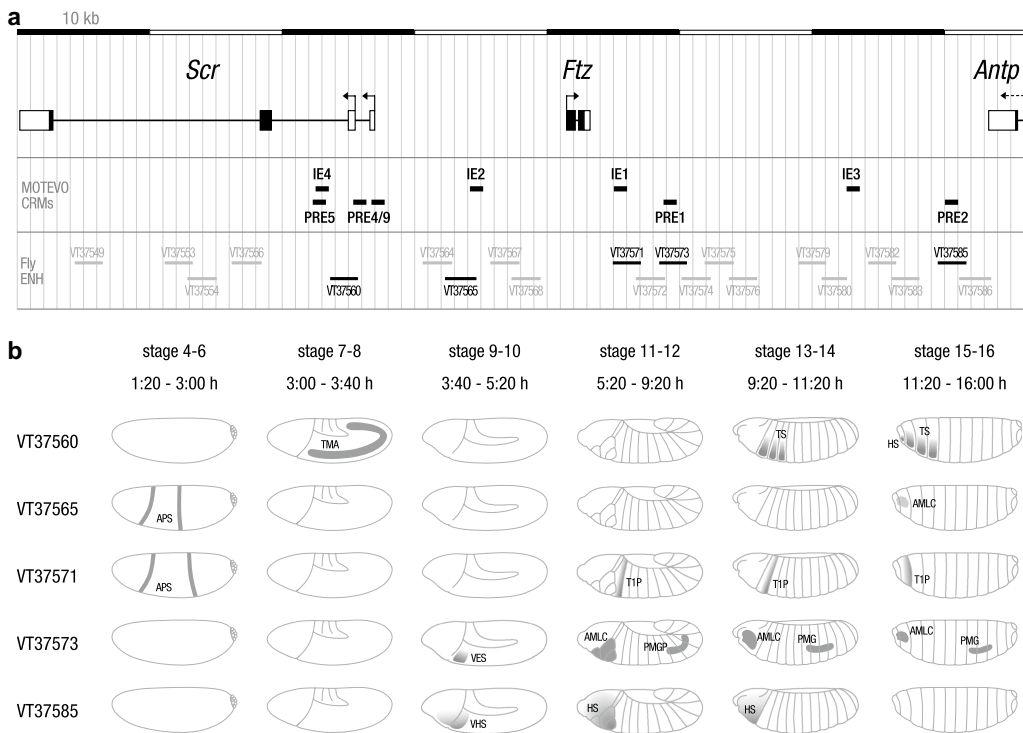


Fig. 5 Enhancers reported by the Fly Enhancers project in the *Scr* region and their expression profiles. **a** The potential enhancers of the studied region reported by the Fly Enhancers project are represented by thick black or gray lines below the diagrams that illustrate gene content and the location of the CRMs identified by MOTEVO (see Fig. 3). Enhancer candidates that overlap, totally or partially, with the MOTEVO CRMs are indicated in black. **b** Expression profiles of the potential enhancers that overlap the MOTEVO CRMs at different

stages of the embryonic development. Gray areas correspond to the embryo regions where the potential enhancer drove the expression of the reporter Gal 4 gene. TMA trunk mesoderm anlage, TS thoracic segments, APS anteroposterior stripes, AMLC antenno-maxillary-labial complex, TIP segment T1 posterior, VES ventral epidermis subset, PMGP posterior midgut primordium, PMG posterior midgut, VHS ventral and head subset, HS head subset

the seven stripes in the *Ftz* expression profile. Nevertheless, both enhancer candidates also drive low expression at later stages in the antenno-maxillary-labial complex and T1 segment, respectively. In contrast, the three enhancers overlapping PRE candidates drive expression mainly at late stages and, with some differences, in regions associated totally or partially with normal *Scr* expression. Thus, VT37560, which overlaps partially with PRE4, begins driving expression at stages 7–8 in the trunk mesoderm but continues directing expression at later stages (13–16) in the three thoracic segments. VT37573, which includes PRE1, drives expression at stage 9–10 in the anterior ventral epidermis and at later stages (11–16) in the antenno-maxillary-labial complex and the posterior midgut. Finally, VT37585, which colocalizes with PRE2, drives expression at stages 9–14 in the head of the embryo.

Finally, there is also a correspondence between some of the IEs and PREs predicted by MOTEVO that do not overlap with any ChIP data frame and the enhancer candidates. Thus, IE9, E6, and IE8 are included in VT37556, VT37564, and VT37582, respectively. Moreover, PRE8 is included in VT37575 and both PRE7 and PRE10 in VT37583. Thus, it remains possible that these CRMs identified by MOTEVO are functional regulatory elements.

Discussion

The strategy developed herein, which combines the compilation of high throughput ChIP data obtained in *D. melanogaster* mainly by the BDTNP and modENCODE projects with the bioinformatic analysis performed using MOTEVO,

allowed the identification of candidate regulatory elements in the *Scr* region that might have a relevant function. In addition, as MOTEVO is based on the interspecific conservation of the elements in the species studied, the detected regulatory elements have been evolutionary conserved at least since the divergence of the *melanogaster* and *obscura* groups that occurred ~24 million years ago (Obbard et al. 2012). In fact, only PRE9 is not shared by all the species studied herein, as this element was only detected in the species of the *melanogaster* subgroup (Online Resource 2).

The present results can be compared with previous data on the *Scr* regulatory region obtained using different approaches summarized in Fig. 6. The most exhaustive functional study of the *Scr* regulatory region to date

was performed in *D. melanogaster* using P-element-mediated transformation with different fragments of this region linked to reporter genes. Along the ~75 kb region between the 3' ends of the *Scr* and *Antp* genes, eight main *Scr* CRMs were identified: 8.2XX, 8.5BR, 2.9BR, 5.5HH, 0.8XH, 3.7HH, 7.0RR, and 10.0XX (Gindhart and Kaufman 1995; Gindhart et al. 1995; Gorman and Kaufman 1995). Prior to this, the central part of the *Scr* regulatory region around *Ftz* had been analyzed using a similar approach with the identification of five regulatory fragments: PS4A, DE, PE, NE, and ZE (Hiromi et al. 1985; Pick et al. 1990; Topol et al. 1991; Han et al. 1993; Schier and Gehring 1993). Three of these elements overlap with one of the previously mentioned *Scr* CRMs.

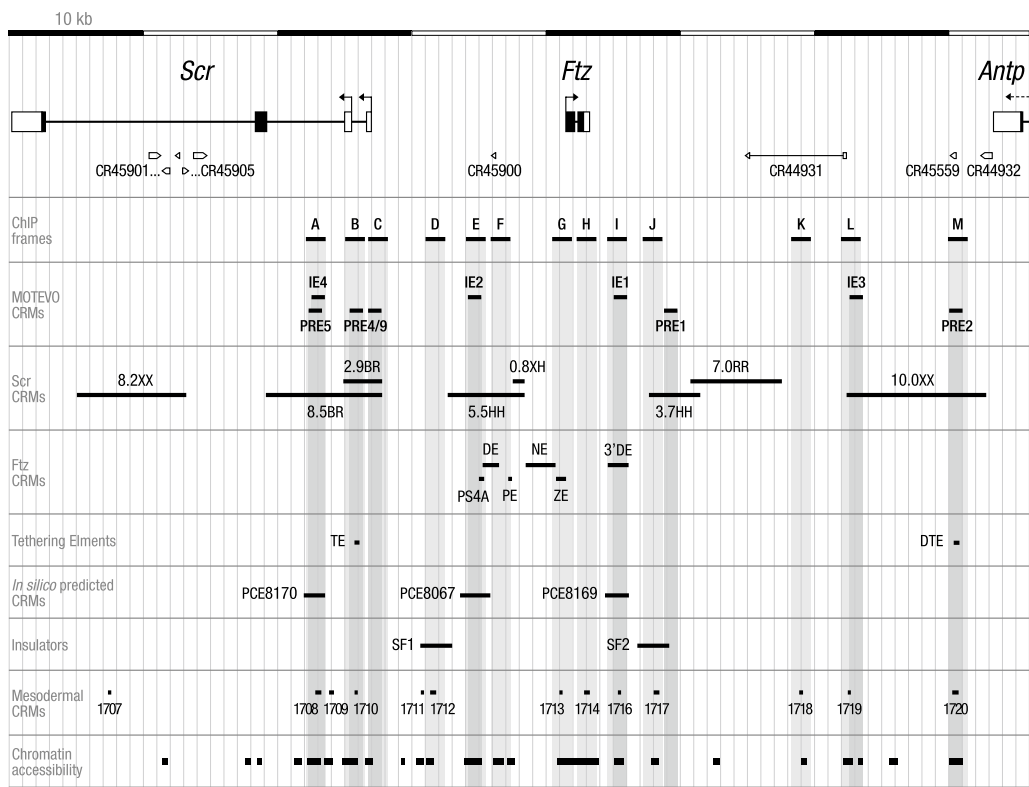


Fig. 6 Summary of the relevant regulatory elements reported herein and elsewhere. The ChIP frames and the main CRMs detected by MOTEVO (IEs and PREs), described in Figs. 1 and 2, are indicated under the gene diagram, which includes the noncoding genes present in the *Scr* regulatory region according to Flybase. ChIP frames, IEs, and PREs are shadowed to provide a positional reference. In the next tracks, the previously published regulatory elements are indicated with thick lines. From top to bottom: *Scr* CRMs, *Ftz* CRMs, tether-

ing elements, in silico-predicted CRMs, insulators, and mesodermal CRMs. Finally, chromatin accessibility data are shown, with thick lines representing DNase I hypersensitive regions. PS4A parasegment 4 activator, DE distal enhancer, PE proximal enhancer, NE neurogenic element, ZE zebra element, 3'DE 3' distal enhancer, TE tethering element, DTE distal tethering element, SF1 and SF2 insulators *Scr*-*Ftz* 1 and 2, respectively. See text for references

Thereafter, the *Scr* regulatory region was analyzed using bioinformatic methods that allowed the detection of an additional *Ftz* CRM (3'DE) and two tethering elements (TE and DTE) apparently involved in the interaction between the proximal and distal CRMs of *Scr* (Calhoun et al. 2002; Calhoun and Levine 2003).

In addition to these studies that focused on the *Scr* regulatory region, genome-wide surveys have also provided information on this region. Thus, in silico genomic searches of clusters of five segmentation factor (*Bcd*, *Cad*, *Hb*, *Kni*, and *Kr*) binding sites (Berman et al. 2004) detected three CRM candidates: PCE8169, which coincides almost exactly with the 3'DE *Ftz* CRM; PCE8170, which is part of the 8.5BR *Scr* CRM; and PCE8067, which colocalizes with the 5.5HH, PS4A, and DE CRMs. On the other hand, a ChIP-chip approximation searching for insulators (Nègre et al. 2010) detected two insulators surrounding *Ftz*: SF1, as previously identified by Belozarov et al. (2003), and SF2, which was subsequently better characterized by Li et al. (2015). Another ChIP-chip experiment, based on the binding profiles of the mesoderm specific transcription factors encoded by *Twist*, *Tinman*, *Myocyte enhancing factor 2*, *Bagpipe*, and *Biniou*, identified 13 candidate mesodermal CRMs in the *Scr* regulatory region (Zinzen et al. 2009). Finally, chromatin accessibility data obtained by DNase I hypersensitivity experiments at stage 5 embryos (Thomas et al. 2011) provide a good support for the presence of the CRMs obtained by ChIP and bioinformatic methods.

In general, the previously characterized *Scr* CRMs show relatively good correspondence with the most relevant CRMs identified herein (Fig. 6). The exceptions are the *Scr* CRMs 8.2XX, 0.8XH, and 7.0RR that do not overlap with any MOTEVO CRMs, or ChIP frames. However, 8.2XX and 7.0RR include, respectively, IE7 and PRE8, which are two CRMs predicted by MOTEVO that do not overlap with any ChIP frame. The present data also help infer a more precise map of the *Scr* regulatory region and to delimit more accurately the different regulatory elements. Thus, the 10.0XX *Scr* CRM actually contains at least two independent CRMs: IE3 and PRE2, which overlap with frames L and M, respectively. Similarly, 2.9BR can be further subdivided into two CRMs: PRE4 and PRE9, which in this case overlap with frames B and C, respectively.

The four most relevant IE candidates (IE1, IE2, IE3, and IE4) identified by MOTEVO overlap with ChIP frames (A, E, I, and L), DNase I hypersensitive regions, and functional regulatory elements previously identified. In fact, IE1 coincides with PCE8169, the mesodermal CRM 1716 and the well-known 3'DE *Ftz* CRM, and thus it is likely to be a *Ftz* regulator. Similarly, IE2 overlaps with 5.5HH, PS4A, PCE8067, and the enhancer candidate VT37565 (Fig. 5), which drives expression in a *Ftz*-like pattern at early stages,

but also in an apparent *Scr*-like pattern later on. Thus, it is difficult to assess whether IE2 regulates the expression of *Scr*, *Ftz*, or even *Antp*, as PS4A also drives expression in a *Antp*-like pattern (Pick et al. 1990). IE3, which had not been identified by bioinformatic methods previously, overlaps with the mesodermal CRM 1719 and with one end of the 10.0XX element. Finally, IE4 overlaps almost totally with PRE5 as well as with the mesodermal CRM 1708 and PCE8170. Thus, IE4 is the only CRM identified by MOTEVO that presents a dual IE/PRE role. This duality is supported by the overlap with the ChIP frame A, in which besides segmentation proteins they also bind some PcG and TrxG proteins, their cofactors and two histone deacetylases.

Besides IEs, five relevant PRE candidates were identified by MOTEVO: PRE1, PRE2, PRE4, PRE5, and PRE9. Although PRE1 is the only PRE that does not colocalize with any ChIP frame or an accessible chromatin region, it has the highest MOTEVO enhancer score and overlaps with both the 3.7HH element and VT37573 (Fig. 5). The 3.7HH element is one of the previously identified enhancers with the strongest evidence for driving *Scr* expression in the prothoracic, labial, and maxillary segments of the embryo (Gindhart et al. 1995; Gorman and Kaufman 1995; Calhoun and Levine 2003), and this is supported by the expression driven by VT37573. Therefore, the PRE1 candidate is probably an enhancer that is rich in PHO and TRL binding sites (Fig. 3) and involved in *Scr* expression, although the ChIP data question whether it is a PRE-type element.

In contrast to PRE1, the already mentioned PRE5 (which overlaps IE4) and the other most relevant PRE candidates (PRE2, PRE4, and PRE9) colocalize with ChIP frames (A, B, C, and M) and DNase I hypersensitive regions. Therefore, they seem to be essential functional PREs. PRE4 and PRE9 are adjacent to the two alternative *Scr* transcription start sites and, thus, they could contribute to the different regulatory requirements of both transcripts. However, only PRE4 colocalizes with a mesodermal CRM, 1710, which may indicate that PRE9 directs expression in nonmesodermal tissues. Finally, PRE2 coincides with the mesodermal CRM 1720 and with the noncoding gene CR45559, which may have functional implications (Hogga and Karch 2002; Herzog et al. 2014; Erokhin et al. 2015). CR45559 is the only noncoding gene in the *Scr* regulatory region that colocalizes with a PRE candidate and, in addition, it is expressed at the same stages of development (modENCODE RNAseq data and Graveley et al. 2011) and in the same tissues (Calhoun and Levine 2003) as *Scr*.

Although we have focused on those CRMs identified by MOTEVO that colocalize with ChIP frames (and PRE1), many of the other IEs and PREs overlap with enhancer candidates tested by the Fly enhancer project or with

previously identified regulatory elements, indicating that they cannot be completely discarded as functional regulatory elements. As an example, IE6 (Fig. 2) could drive expression in anteroposterior stripes according to its colocalization with VT37564. It also overlaps with a DNase I hypersensitive region and with the mesodermal enhancer 1711.

The ChIP data analysis also allowed the better characterization of the two previously identified insulator elements: SF1 and SF2. Data indicate that both insulators recruit the Polycomb protein PHO and a few Trithorax proteins. As far as we know, there are no previous data relating PHO to either insulator elements or their associated proteins. In contrast to PHO, the TrxG protein NEJ has been found to be associated with insulators (Philip et al. 2015), but it was not detected in this study either in SF1 or SF2. In addition, the present ChIP data analysis revealed clear binding of the insulator-associated protein MOD (MDG4) at PRE2, PRE4 and PRE9. Melnikova et al. (2004) detected a direct interaction between MOD (MDG4) and the PcG cofactor TRL that mediated the bypassing of the insulator elements to allow the interaction between a proximal and a distal CRM. The compartmentalization of different regulatory regions by insulators and the PcG/TrxG-mediated regulation are two processes that involve similar mechanisms of chromatin remodeling with the formation of 3D structures (Bantignies and Cavalli 2011; Matzat and Lei 2014). It has previously been observed that these two processes affect each other (Van Bortle et al. 2012; Li et al. 2013) and, therefore, it is not unexpected that insulators and PREs share some of their binding proteins, such as PHO or MOD (MDG4).

The *Scr* regulatory region also includes two tethering elements identified by Calhoun and Levine (2003). The authors suggested that a common set of proteins would bind to these elements and form a chromatin loop that would facilitate the interaction between the *Scr* promoter and the distal regulatory elements of the gene, such as the 3.7HH element. It is notable that the two tethering elements are actually part of the PRE2 and PRE4 elements identified herein and, therefore, PcG and TrxG proteins and their cofactors probably help to form the chromatin loop. However, no PcG, TrxG, or insulator protein binds to the TTC GAA motif that characterizes the tethering elements and, as far as we know, the protein that recognizes this motif has not yet been identified. In addition, we have not found tethering elements associated with other PRE candidates along the Ant-C and Bx-C regions, so they do not seem to be a common feature of PREs.

The last candidate regulatory element detected in the ChIP data analysis is frame K. This element, which was first identified here, does not colocalize with any MOTEVO peak and has a protein binding profile that cannot be clearly associated with an IE, PRE, or insulator.

In fact, frame K binds only to PHO, PH, and four histone deacetylases. Three of these deacetylases (HDAC1, HDAC4, and HDAC6) are also common in PRE2, PRE4, PRE5, and PRE9, but HDAC11 does not bind to any other ChIP frame. This frame also coincides with the mesodermal CRM 1718 (Fig. 6), a DNase I hypersensitive region and the enhancer candidate VT37579 (Fig. 5), which drives expression in the head, brain, and anteroventral nerve cord. Further experimental analysis is required to ascertain whether this element is relevant in *Scr* regulation.

Finally, the observed differences in the density of transcription factor binding sites in the CRMs, especially those detected between PHO and TRL (Fig. 3), are concordant with available information. Thus, all identified PREs have at least one PHO binding site, but none of them has more than two. Schuettengruber et al. (2014) observed that PHO–DNA interactions occur over a wide spectrum of noncanonical, low-affinity binding sites, indicating that PHO recruitment relies mainly on cooperative effects with other transcription factors rather than on the presence of binding sites. One of these transcription factors is in fact TRL (Mahmoudi et al. 2003). In contrast with PHO, a high saturation of canonical TRL binding sites is present in PREs, which is consistent with previous studies (Horard et al. 2000; Cunningham et al. 2010). TRL binding is largely self-dependent and does not require secondary transcription factors (Cheng et al. 2013), which is consistent with the proposal that TRL is a pioneer factor (Zaret and Carroll 2011). These differences between factors, which were corroborated in the present study, suggest a model in which the auto-dependent TRL paves the way for the subsequent binding of PHO helped by cooperative interactions with other PcG proteins.

In summary, the present analysis contributes to a better understanding of the relevant functional elements of the *Scr* regulatory region. In fact, in addition to confirming and delimiting more accurately the previously detected CRMs, it has allowed the identification of a candidate element with a dual IE/PRE function (IE4/PRE5), two adjacent PREs (PRE4 and PRE 9) that may control the alternative transcription start sites of *Scr*, a PRE (PRE2) overlapping with a noncoding gene, and a ChIP frame (K) that is highly enriched in histone deacetylases. Furthermore, the regulatory proteins that bind to the two insulator elements have been properly characterized and it has been shown that the two tethering elements (TE and DTE) are part of two PREs (PRE4 and PRE2, respectively). Therefore, the present results could assist with the design of future experimental approaches to gain new insights into the expression control of the *Scr* region and to better define the regulatory region of other developmental genes.

Acknowledgements We thank Montserrat Aguadé for critical comments on the manuscript. This study was funded by a predoctoral fellowship from the Agència de Gestió d'Ajuts Universitaris i de Recerca (AGAUR), Generalitat de Catalunya, Catalonia, Spain, to JMC-M; and grants BFU2012-35168 and BFU2015-63732 from the Ministerio de Economía y Competitividad, Spain, and 2009SGR-1287 and 2014SGR-10555 from the Comissió Interdepartamental de Recerca i Innovació Tecnològica, Catalonia, Spain, to M Aguadé.

Compliance with ethical standards

Conflict of interest The authors declare that there is no conflict of interest.

Ethical approval This article does not contain any studies with human participants or animals performed by the authors.

References

- Americo J, Whiteley M, Lesley Brown J, Fujioka M, Jaynes JB, Kassiss JA (2002) A complex array of DNA-binding proteins required for pairing-sensitive silencing by a polycomb group response element from the *Drosophila* engrailed gene. *Genetics* 160:1561–1571
- Andrew DJ (1995) The Sex combs reduced gene of *Drosophila melanogaster* has multiple transcripts. *Gene* 152:149–155
- Arnold P, Erb I, Pachkov M, Molina N, Van Nimwegen E (2012) MotEvo: integrated bayesian probabilistic methods for inferring regulatory sites and motifs on multiple alignments of DNA sequences. *Bioinformatics* 28:487–494
- Bantignies F, Cavalli G (2011) Polycomb group proteins: repression in 3D. *Trends Genet* 27:454–464
- Belozzerov VE, Majumder P, Shen P, Cai HN (2003) A novel boundary element may facilitate independent gene regulation in the Antennapedia complex of *Drosophila*. *EMBO J* 22:3113–3121
- Berman BP, Nibu Y, Pfeiffer BD, Tomancak P, Celniker SE, Levine M, Rubin GM, Eisen MB (2002) Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the *Drosophila* genome. *Proc Natl Acad Sci USA* 99:757–762
- Berman BP, Pfeiffer BD, Laverty TR, Salzberg SL, Rubin GM, Eisen MB, Celniker SE (2004) Computational identification of developmental enhancers: conservation and function of transcription factor binding-site clusters in *Drosophila melanogaster* and *Drosophila pseudoobscura*. *Genome Biol* 5:R61
- Blastyák A, Mishra RK, Karch F (2006) Efficient and specific targeting of polycomb group proteins requires cooperative interaction between grainyhead and pleiohomeotic. *Mol Cell Biol* 26:1434–1444
- Brown JL, Kassiss JA (2010) Spps, a *Drosophila* Sp1/KLF family member, binds to PREs and is required for PRE activity late in development. *Development* 137:2597–2602
- Brown JL, Kassiss JA (2013) Architectural and functional diversity of polycomb group response elements in *Drosophila*. *Genetics* 195:407–419
- Brown JL, Mucci D, Whiteley M, Dirksen ML, Kassiss JA (1998) The *Drosophila* polycomb group gene pleiohomeotic encodes a DNA binding protein with homology to the transcription factor YY1. *Mol Cell* 1:1057–1064
- Brown JL, Fritsch C, Mueller J, Kassiss JA (2003) The *Drosophila* pho-like gene encodes a YY1-related DNA binding protein that is redundant with pleiohomeotic in homeotic gene silencing. *Development* 130:285–294
- Calhoun VC, Levine M (2003) Long-range enhancer-promoter interactions in the Scr-Antp interval of the *Drosophila* Antennapedia complex. *Proc Natl Acad Sci USA* 100:9878–9883
- Calhoun VC, Stathopoulos A, Levine M (2002) Promoter-proximal tethering elements regulate enhancer-promoter specificity in the *Drosophila* Antennapedia complex. *Proc Natl Acad Sci USA* 99:9243–9247
- Celniker SE, Dillon LAL, Gerstein MB, Gunsalus KC, Henikoff S, Karpen GH, Kellis M, Lai EC, Lieb JD, MacAlpine DM, Micklem G, Piano F, Snyder M, Stein L, White KP, Waterston RH (2009) Unlocking the secrets of the genome. *Nature* 459:927–930
- Cheng Q, Kazemian M, Pham H, Blatti C, Celniker SE, Wolfe SA, Brodsky MH, Sinha S (2013) Computational identification of diverse mechanisms underlying transcription factor-DNA occupancy. *PLoS Genet* 9:e1003571
- Cunningham MD, Brown JL, Kassiss JA (2010) Characterization of the polycomb group response elements of the *Drosophila melanogaster* invected Locus. *Mol Cell Biol* 30:820–828
- Déjardin J, Rappailles A, Cuvier O, Grimaud C, Decoville M, Locker D, Cavalli G (2005) Recruitment of *Drosophila* Polycomb group proteins to chromatin by DSP1. *Nature* 434:533–538
- Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 32:1792–1797
- Erokhin M, Elizar'ev P, Parshikov A, Schedl P, Georgiev P, Chetverina D (2015) Transcriptional read-through is not sufficient to induce an epigenetic switch in the silencing activity of Polycomb response elements. *Proc Natl Acad Sci USA* 112:14930–14935
- Fiedler T, Rehmsmeier M (2006) jPREdictor: a versatile tool for the prediction of cis-regulatory elements. *Nucleic Acids Res* 34:546–550
- Fritsch C, Brown JL, Kassiss JA, Müller J (1999) The DNA-binding polycomb group protein pleiohomeotic mediates silencing of a *Drosophila* homeotic gene. *Development* 126:3905–3913
- Gindhart JG, Kaufman TC (1995) Identification of Polycomb and trithorax group responsive elements in the regulatory region of the *Drosophila* homeotic gene sex combs reduced. *Genetics* 139:797–814
- Gindhart JG, King AN, Kaufman TC (1995) Characterization of the regulatory region of the *Drosophila* homeotic gene sex combs reduced. *Genetics* 139:781–795
- Gorman MJ, Kaufman TC (1995) Genetic analysis of embryonic cis-acting regulatory elements of the *Drosophila* homeotic gene sex combs reduced. *Genetics* 140:557–572
- Graveley BR, Brooks AN, Carlson JW, Duff MO, Landolin JM, Yang L, Artieri CG, van Baren MJ, Boley N, Booth BW, Brown JB, Chervas L, Davis CA, Dobin A, Li R, Lin W, Malone JH, Mattiuzzo NR, Miller D, Sturgill D, Tuch BB, Zaleski C, Zhang D, Blanchette M, Dudoit S, Eads B, Green RE, Hammonds A, Jiang L, Kapranov P, Langton L, Perrimon N, Sandler JE, Wan KH, Willingham A, Zhang Y, Zou Y, Andrews J, Bickel PJ, Brenner SE, Brent MR, Chervas P, Gingeras TR, Hoskins RA, Kaufman TC, Oliver B, Celniker SE (2011) The developmental transcriptome of *Drosophila melanogaster*. *Nature* 471:473–479
- Han W, Yu Y, Altan N, Pick L (1993) Multiple proteins interact with the fushi tarazu proximal enhancer. *Mol Cell Biol* 13:5549–5559
- Hauenschild A, Ringrose L, Altmutter C, Paro R, Rehmsmeier M (2008) Evolutionary plasticity of polycomb/trithorax response elements in *Drosophila* species. *PLoS Biol* 6:e261
- Herzog VA, Lempradl A, Trupke J, Okulski H, Altmutter C, Ruge F, Boidol B, Kubicek S, Schmauss G, Aumayr K, Ruf M, Pospisilik A, Dimond A, Senegin HB, Vargas ML, Simon JA, Ringrose L (2014) A strand-specific switch in noncoding transcription switches the function of a Polycomb/Trithorax response element. *Nat Genet* 46:973–981

- Hiromi Y, Kuroiwa A, Gehring WJ (1985) Control elements of the *Drosophila* segmentation gene fushi tarazu. *Cell* 43:603–613
- Hogga I, Karch F (2002) Transcription through the iab-7 cis-regulatory domain of the bithorax complex interferes with maintenance of Polycomb-mediated silencing. *Development* 129:4915–4922
- Horard B, Tatout C, Poux S, Pirrotta V (2000) Structure of a polycomb response element and in vitro binding of polycomb group complexes containing GAGA factor. *Mol Cell Biol* 20:3187–3197
- Hughes CL, Kaufman TC (2002) Hox genes and the evolution of the arthropod body plan. *Evol Dev* 4:459–499
- Hur M-W, Laney JD, Jeon SH, Ali J, Biggin MD (2002) Zeste maintains repression of Ubx transgenes: support for a new model of Polycomb repression. *Development* 129:1339–1343
- Isshiki T, Pearson B, Holbrook S, Doe CQ (2001) *Drosophila* neuroblasts sequentially express transcription factors which specify the temporal identity of their neuronal progeny. *Cell* 106:511–521
- Kassis JA, Brown JL (2013) Polycomb group response elements in *Drosophila* and vertebrates. *Adv Genet* 81:83–118
- Kaufman TC, Lewis R, Wakimoto B (1980) Cytogenetic analysis of chromosome 3 in *Drosophila melanogaster*: the homeotic gene complex in polytene chromosome interval 84 A-B. *Genetics* 94:115–133
- Kaufman TC, Seeger MA, Olsen G (1990) Molecular and genetic organization of the antennapedia gene complex of *Drosophila melanogaster*. *Adv Genet* 27:309–362
- Kuroiwa A, Kloter U, Baumgartner P, Gehring WJ (1985) Cloning of the homeotic Sex combs reduced gene in *Drosophila* and in situ localization of its transcripts. *EMBO J* 4:3757–3764
- Kvon EZ, Kazmar T, Stampfel G, Yáñez-Cuna JO, Pagani M, Schernhuber K, Dickson BJ, Stark A (2014) Genome-scale functional characterization of *Drosophila* developmental enhancers in vivo. *Nature* 512:91–95
- Lanzuolo C, Orlando V (2012) Memories from the polycomb group proteins. *Annu Rev Genet* 46:561–589
- Lemotte PK, Kuroiwa A, Fessler L, Gehring WJ (1989) The homeotic gene sex combs reduced of *Drosophila*: gene structure and embryonic expression. *EMBO J* 8:219–227
- Li X, MacArthur S, Bourgon R, Nix D, Pollard DA, Iyer VN, Hechmer A, Simirenko L, Stapleton M, Luengo Hendriks CL, Chu HC, Ogawa N, Inwood W, Sementchenko V, Beaton A, Weiszmarm R, Celniker SE, Knowles DW, Gingeras T, Speed TP, Eisen MB, Biggin MD (2008) Transcription factors bind thousands of active and inactive regions in the *Drosophila* blastoderm. *PLoS Biol* 6:e27
- Li HB, Ohno K, Gui H, Pirrotta V (2013) Insulators target active genes to transcription factories and polycomb-repressed genes to polycomb bodies. *PLoS Genet* 9:e1003436
- Li M, Ma Z, Liu JK, Roy S, Patel SK, Lane DC, Cai HN (2015) An organizational hub of developmentally regulated chromatin loops in the *Drosophila* Antennapedia complex. *Mol Cell Biol* 35:4018–4029
- MacArthur S, Li X-Y, Li J, Brown JB, Chu HC, Zeng L, Grondona BP, Hechmer A, Simirenko L, Keränen SVE, Knowles DW, Stapleton M, Bickel P, Biggin MD, Eisen MB (2009) Developmental roles of 21 *Drosophila* transcription factors are determined by quantitative differences in binding to an overlapping set of thousands of genomic regions. *Genome Biol* 10:R80
- Maeda RK, Karch F (2006) The ABC of the BX-C: the bithorax complex explained. *Development* 133:1413–1422
- Mahaffey JW, Kaufman TC (1987) Distribution of the Sex combs reduced gene products in *Drosophila melanogaster*. *Genetics* 117:51–60
- Mahmoudi T, Zuijderduijn LMP, Mohd-Sarip A, Verrijzer CP (2003) GAGA facilities binding of pleiohomeotic to a chromatinized Polycomb response element. *Nucleic Acids Res* 31:4147–4156
- Matzat LH, Lei EP (2014) Surviving an identity crisis: a revised view of chromatin insulators in the genomics era. *Biochim Biophys Acta* 1839:203–214
- Melnikova L, Juge F, Gruzdeva N, Mazur A, Cavalli G, Georgiev P (2004) Interaction between the GAGA factor and Mod(mdg4) proteins promotes insulator bypass in *Drosophila*. *Proc Natl Acad Sci USA* 101:14806–14811
- Müller J, Verrijzer P (2009) Biochemical mechanisms of gene regulation by polycomb group protein complexes. *Curr Opin Genet Dev* 19:150–158
- Nègre N, Brown CD, Shah PK, Kheradpour P, Morrison CA, Henikoff JG, Feng X, Ahmad K, Russell S, White RAH, Stein L, Henikoff S, Kellis M, White KP (2010) A comprehensive map of insulator elements for the *Drosophila* genome. *PLoS Genet* 6:e1000814
- Nègre N, Brown CD, Ma L, Bristow CA, Miller SW, Wagner U, Kheradpour P, Eaton ML, Loriaux P, Sealfon R, Li Z, Ishii H, Spokony RF, Chen J, Hwang L, Cheng C, Auburn RP, Davis MB, Domanus M, Shah PK, Morrison CA, Zieba J, Suchy S, Senderowicz L, Victorson A, Bild NA, Grundstad AJ, Hanley D, MacAlpine DM, Mannervik M, Venken K, Bellen H, White R, Gerstein M, Russell S, Grossman RL, Ren B, Posakony JW, Kellis M, White KP (2011) A cis-regulatory map of the *Drosophila* genome. *Nature* 471:527–531
- Obbard DJ, MacLennan J, Kim K-W, Rambaut A, O'Grady PM, Jiggins FM (2012) Estimating divergence dates and substitution rates in the *Drosophila* phylogeny. *Mol Biol Evol* 29:3459–3473
- Okulski H, Druck B, Bhalarao S, Ringrose L (2011) Quantitative analysis of polycomb response elements (PREs) at identical genomic locations distinguishes contributions of PRE sequence and genomic environment. *Epigenetics Chromatin* 4:4
- Orsi GA, Kasinathan S, Hughes KT, Saminadin-Peter S, Henikoff S, Ahmad K (2014) High-resolution mapping defines the cooperative architecture of Polycomb response elements. *Genome Res* 24:809–820
- Peel AD, Chipman AD, Akam M (2005) Arthropod segmentation: beyond the *Drosophila* paradigm. *Nat Rev Genet* 6:905–916
- Philip P, Boija A, Vaid R, Churcher AM, Meyers DJ, Cole PA, Mannervik M, Stenberg P (2015) CBP binding outside of promoters and enhancers in *Drosophila melanogaster*. *Epigenetics Chromatin* 8:48
- Pick L, Heffer A (2012) Hox gene evolution: multiple mechanisms contributing to evolutionary novelties. *Ann N Y Acad Sci* 1256:15–32
- Pick L, Schier A, Affolter M, Schmidt-Glenewinkel T, Gehring WJ (1990) Analysis of the ftz upstream element: germ layer-specific enhancers are independently autoregulated. *Genes Dev* 4:1224–1239
- Reuter R, Scott MP (1990) Expression and function of the homeotic genes Antennapedia and Sex combs reduced in the embryonic midgut of *Drosophila*. *Development* 109:289–303
- Riley PD, Carrol SB, Scott MP (1987) The expression and regulation of sex combs reduced protein in *Drosophila* embryos. *Genes Dev* 5:716–730
- Ringrose L, Rehmsmeier M, Dura JM, Paro R (2003) Genome-wide prediction of polycomb/trithorax response elements in *Drosophila melanogaster*. *Dev Cell* 5:759–771
- Roy S, Ernst J, Kharchenko PV, Kheradpour P, Negre N, Eaton ML, Landolin JM, Bristow CA, Ma L, Lin MF, Washietl S, Arshinoff BI, Ay F, Meyer PE, Robine N, Washington NL, Di Stefano L, Berezhikov E, Brown CD, Candeias R, Carlson JW, Carr A, Jungreis I, Marbach D, Sealfon R, Tolstorukov MY, Will S, Alekseyenko AA, Artieri C, Booth BW, Brooks AN, Dai Q, Davis CA, Duff MO, Feng X, Gorchakov AA, Gu T, Henikoff JG, Kapranov P, Li R, MacAlpine HK, Malone J, Minoda A,

- Nordman J, Okamura K, Perry M, Powell SK, Riddle NC, Sakai A, Samsonova A, Sandler JE, Schwartz YB, Sher N, Spokony R, Sturgill D, van Baren M, Wan KH, Yang L, Yu C, Feingold E, Good P, Guyer M, Lowdon R, Ahmad K, Andrews J, Berger B, Brenner SE, Brent MR, Cherbas L, Elgin SCR, Gingeras TR, Grossman R, Hoskins RA, Kaufman TC, Kent W, Kuroda MI, Orr-Weaver T, Perrimon N, Pirrotta V, Posakony JW, Ren B, Russell S, Cherbas P, Graveley BR, Lewis S, Micklem G, Oliver B, Park PJ, Celniker SE, Henikoff S, Karpen GH, Lai EC, MacAlpine DM, Stein LD, White KP, Kellis M (2010) Identification of functional elements and regulatory circuits by *Drosophila* modENCODE. *Science* 330:1787–1797
- Schier AF, Gehring WJ (1993) Analysis of a fushi tarazu autoregulatory element: multiple sequence elements contribute to enhancer activity. *EMBO J* 12:1111–1119
- Schroeder MD, Pearce M, Fak J, Fan H, Unnerstall U, Emberly E, Rajewsky N, Siggia ED, Gaul U (2004) Transcriptional control in the segmentation gene network of *Drosophila*. *PLoS Biol* 2:e271
- Schuettengruber B, Ganapathi M, Leblanc B, Portoso M, Jaschek R, Tolhuis B, Van Lohuizen M, Tanay A, Cavalli G (2009) Functional anatomy of polycomb and trithorax chromatin landscapes in *Drosophila* embryos. *PLoS Biol* 7:e1000013
- Schuettengruber B, Martinez AM, Iovino N, Cavalli G (2011) Trithorax group proteins: switching genes on and keeping them active. *Nat Rev Mol Cell Biol* 12:799–814
- Schuettengruber B, Oded Elkayam N, Sexton T, Entrevan M, Stern S, Thomas A, Yaffe E, Parrinello H, Tanay A, Cavalli G (2014) Cooperativity, specificity, and evolutionary stability of polycomb targeting in *Drosophila*. *Cell Rep* 9:219–233
- Starr MO, Ho MCW, Gunther EJM, Tu Y, Shur AS, Goetz SE, Borok MJ, Kang V, Drewell RA (2012) Molecular dissection of cis-regulatory modules at the *Drosophila* bithorax complex reveals critical transcription factor signature motifs. *Dev Biol* 359:290–302
- Steffen PA, Ringrose L (2014) What are memories made of? How Polycomb and Trithorax proteins mediate epigenetic memory. *Nat Rev Mol Cell Biol* 15:340–356
- Tamura K, Stecher G, Peterson D, Filipski A, Kumar S (2013) MEGA6: Molecular evolutionary genetics analysis version 6.0. *Mol Biol Evol* 30:2725–2729
- Thomas S, Li X-Y, Sabo PJ, Sandstrom R, Thurman RE, Canfield TK, Giste E, Fisher W, Hammonds A, Celniker SE, Biggin MD, Stamatoyannopoulos JA (2011) Dynamic reprogramming of chromatin accessibility during *Drosophila* embryo development. *Genome Biol* 12:R43
- Topol J, Dearolf CR, Prakash K, Parker CS (1991) Synthetic oligonucleotides recreate *Drosophila* fushi tarazu zebra-stripe expression. *Genes Dev* 5:855–867
- Van Bortle K, Ramos E, Takenaka N, Yang J, Wahi JE, Corces VG (2012) *Drosophila* CTCF tandemly aligns with other insulator proteins at the borders of H3K27me3 domains. *Genome Res* 22:2176–2187
- Wang L, Brown JL, Cao R, Zhang Y, Kassiss JA, Jones RS (2004) Hierarchical recruitment of polycomb group silencing complexes. *Mol Cell* 14:637–646
- Zaret KS, Carroll JS (2011) Pioneer transcription factors: establishing competence for gene expression. *Genes Dev* 25:2227–2241
- Zeng J, Kirk BD, Gou Y, Wang Q, Ma J (2012) Genome-wide polycomb target gene prediction in *Drosophila melanogaster*. *Nucleic Acids Res* 40:5848–5863
- Zinzen RP, Girardot C, Gagneur J, Braun M, Furlong EEM (2009) Combinatorial binding predicts spatio-temporal cis-regulatory activity. *Nature* 462:65–70

Sex combs reduced (Scr) regulatory region of *Drosophila* revisited

Calvo-Martín JM, Papaceit M, Segarra C

2017. *Molecular Genetics and Genomics*. doi: 10.1007/s00438-017-1309-1

Supplementary Material

Online Resource 1. Location of the CRMs identified in the present study and of the regulatory elements previously described.

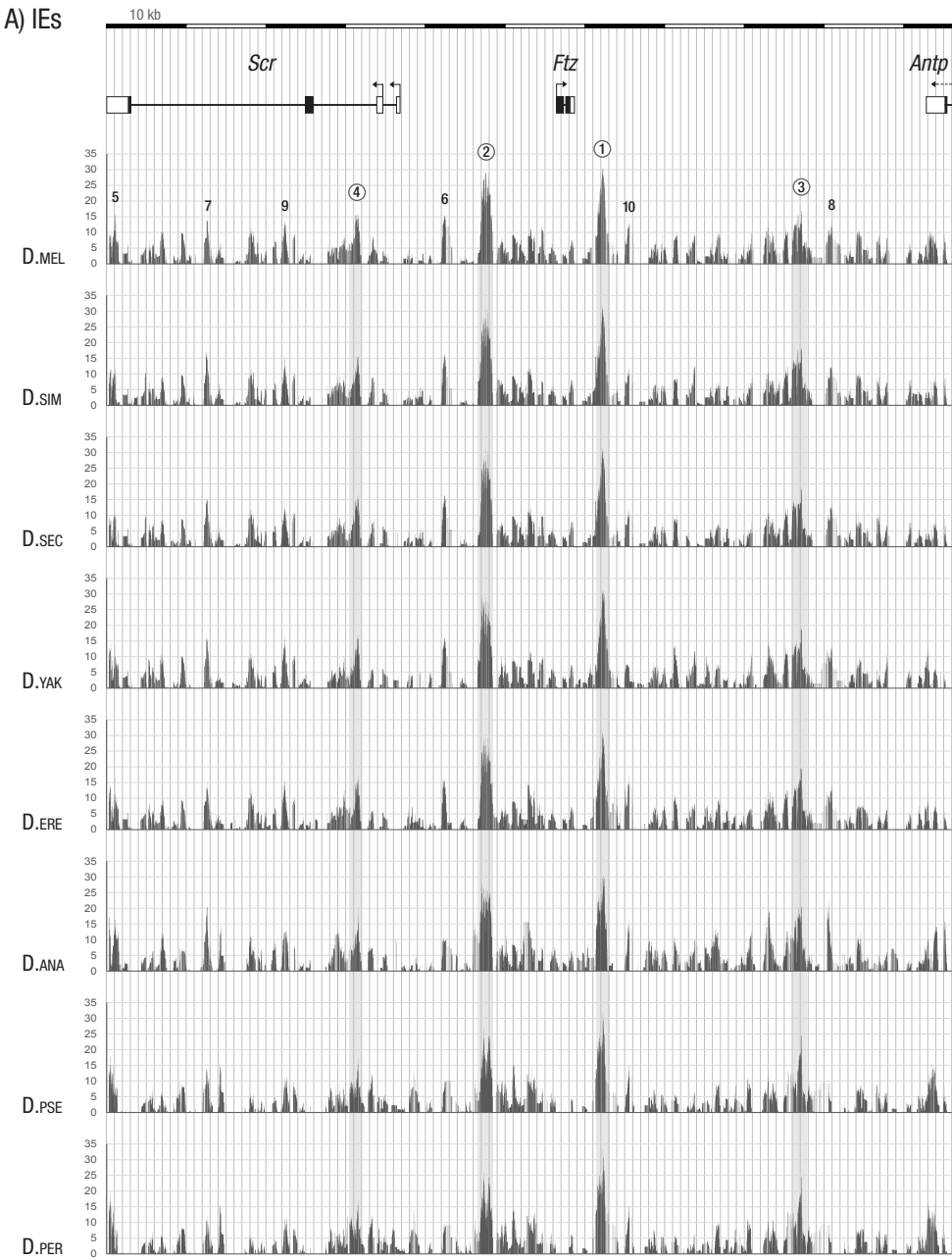
	<i>D. melanogaster</i> release 5 ^a		<i>D. melanogaster</i> release 6.09 ^a		<i>D. melanogaster</i> <i>Scr</i> regulatory region analyzed ^a		Multiple alignment <i>Scr</i> regulatory region analyzed ^a	
MOTEO inferred CRMs								
IE1	2693705	2694705	6867983	6868983	45020	46020	61407	63315
IE2	2682865	2683865	6857143	6858143	34180	35180	46787	48049
IE3	2711295	2712295	6885573	6886573	62610	63610	86280	87536
IE4	2671195	2672195	6845473	6846473	22510	23510	30694	32015
IE5	2649085	2650085	6823363	6824363	400	1400	483	1689
IE6	2679195	2680195	6853473	6854473	30510	31510	41641	43426
IE7	2657175	2658175	6831453	6832453	8490	9490	11891	13271
IE8	2713225	2714225	6887503	6888503	64540	65540	89474	92080
IE9	2664425	2665425	6838703	6839703	15740	16740	21689	22897
IE10	2695225	2696225	6869503	6870503	46540	47540	64403	66337
PRE1	2697455	2698445	6871733	6872723	48770	49770	68247	69482
PRE2	2718695	2719685	6892973	6893963	70010	71010	98554	100007
PRE3	2689965	2690965	6864243	6865243	41280	42280	56382	57585
PRE4	2674025	2675015	6848303	6849293	25340	26340	34526	35934
PRE5	2670975	2671965	6845253	6846243	22290	23290	30376	31675
PRE6	2691015	2692015	6865293	6866293	42330	43330	57642	58923
PRE7	2714885	2715885	6889163	6890163	66200	67200	93185	94458
PRE8	2700065	2701065	6874343	6875343	51380	52380	71414	72649
PRE9	2675395	2676385	6849673	6850663	26710	27710	36559	37991
PRE10	2716255	2717255	6890533	6891533	67570	68570	94903	96257
Scr previously indentified CRMs								
8.2X	2653766	2661845	6828044	6836123	5081	13160	6744	18159
8.5BR	2667897	2676484	6842175	6850762	19212	27799	26023	38091
2.9BR	2673638	2676484	6847916	6850762	24953	27799	33957	38091
5.5H	2681361	2687012	6855639	6861290	32676	38327	44965	52456
0.8XH	2686196	2687012	6860474	6861290	37511	38327	51363	52456
3.7H	2696328	2700152	6870606	6874430	47643	51467	66462	71525
7.0R	2699431	2706220	6873709	6880498	50746	57535	70681	79487
10.0X	2710986	2721405	6885264	6895683	62301	72720	85898	102158
Ftz previously indentified CRMs								
PS4A	2683634	2684045	6857912	6858323	34949	35360	47757	48282
Distal enhancer	2683960	2685130	6858238	6859408	35275	36445	48165	49865
Proximal enh. AE1	2685794	2686116	6860072	6860394	37109	37431	50820	51216
Neurogenic enh.	2687127	2689375	6861405	6863653	38442	40690	52574	55528
Zebra element	2689377	2689659	6863655	6863937	40692	40974	55530	55897
3' DE	2693516	2694767	6867794	6869045	44831	46082	61178	63462
Berman et al. (2004) CRMs								
PCE8170	2670657	2672241	6844935	6846519	21972	23556	29730	32066
PCE8067	2682313	2684590	6856591	6858868	33628	35905	46120	49256
PCE8169	2693266	2694696	6867544	6868974	44581	46011	60878	63305

^a Numbers indicate the exact genomic positions according to: the *D. melanogaster* genome (either release 5 or 6.09), the *Scr* regulatory region analyzed of *D. melanogaster* and the multiple alignment including the *Scr* regulatory region of 8 *Drosophila* species.

^b SF2 position is inferred approximately.

Online Resource 1. continued from previous page.

	<i>D. melanogaster</i> release 5 ^a		<i>D. melanogaster</i> release 6.09 ^a		<i>D. melanogaster</i> Scr regulatory region analyzed ^a		Multiple alignment Scr regulatory region analyzed ^a	
Tethering elements								
TE	2674398	2674767	6848676	6849045	25713	26082	35122	35595
DTE	2718976	2719364	6893254	6893642	70291	70679	98926	99453
Insulators								
SF1	2679453	2681789	6853731	6856067	30768	33104	41976	45439
SF2 ^b	2695685	2697685	6869963	6871963	47000	49000	65218	68547
Fly Enhancer Candidates								
VT37549	2653040	2655142	6827318	6829420	4355	6457	5753	8867
VT37553	2659847	2661969	6834125	6836247	11162	13284	15528	18342
VT37554	2661531	2663752	6835809	6838030	12846	15067	17666	20754
VT37556	2664895	2667156	6839173	6841434	16210	18471	22279	25213
VT37560	2672306	2674395	6846584	6848673	23621	25710	32140	35088
VT37564	2679309	2681544	6853587	6855822	30624	32859	41815	45161
VT37565	2680992	2683269	6855270	6857547	32307	34584	44467	47282
VT37567	2684413	2686617	6858691	6860895	35728	37932	48887	51877
VT37568	2686006	2688195	6860284	6862473	37321	39510	51064	53866
VT37571	2693633	2695762	6867911	6870040	44948	47077	61318	65338
VT37572	2695379	2697581	6869657	6871859	46694	48896	64825	68398
VT37573	2697142	2699236	6871420	6873514	48457	50551	67860	70439
VT37574	2698826	2701039	6873104	6875317	50141	52354	69912	72623
VT37575	2700554	2702776	6874832	6877054	51869	54091	72019	75068
VT37576	2702394	2704501	6876672	6878779	53709	55816	74602	77320
VT37579	2707640	2709759	6881918	6884037	58955	61074	81563	84324
VT37580	2709408	2711499	6883686	6885777	60723	62814	83956	86533
VT37582	2712907	2715012	6887185	6889290	64222	66327	88341	93337
VT37583	2714662	2716763	6888940	6891041	65977	68078	92923	95636
VT37585	2718167	2720313	6892445	6894591	69482	71628	97438	100786
VT37586	2719754	2722170	6894032	6896448	71069	73485	100121	103095
Mesodermal CRMs								
MesoCRM-1707	2656088	2656288	6830366	6830566	7403	7603	10044	10292
MesoCRM-1708	2671476	2671925	6845754	6846203	22791	23240	31101	31620
MesoCRM-1709	2672512	2672817	6846790	6847095	23827	24132	32443	32887
MesoCRM-1710	2674428	2674628	6848706	6848906	25743	25943	35157	35384
MesoCRM-1711	2679348	2679548	6853626	6853826	30663	30863	41857	42077
MesoCRM-1712	2680046	2680440	6854324	6854718	31361	31755	43242	43767
MesoCRM-1713	2689658	2689858	6863936	6864136	40973	41173	55896	56224
MesoCRM-1714	2691462	2691662	6865740	6865940	42777	42977	58158	58453
MesoCRM-1715	2691663	2691863	6865941	6866141	42978	43178	58454	58707
MesoCRM-1716	2693998	2694198	6868276	6868476	45313	45513	61744	62006
MesoCRM-1717	2696654	2697045	6870932	6871323	47969	48360	67175	67746
MesoCRM-1718	2707488	2707748	6881766	6882026	58803	59063	81220	81768
MesoCRM-1719	2711043	2711243	6885321	6885521	62358	62558	85991	86226
MesoCRM-1720	2718856	2719271	6893134	6893549	70171	70586	98778	99356



Online Resource 2. Bioinformatic detection of CRMs along the *Scr* regulatory region in 8 *Drosophila* species. The gene content of the studied region is shown in the top diagrams (see figure 1). Graphs below show the enhancer scores (y-axis) provided by MOTEVO along the *Scr* region (x-axis) for each one of the 8 *Drosophila* species analyzed. In contrast with Figure 2, nucleotide positions do not correspond to the *D. melanogaster* sequence, but to the multiple alignment. Shadows link the main peaks in *D. melanogaster* with the equivalent peaks in the other species. **a)** IE predictions using a set of TFBSs for segmentation regulatory proteins. **b)** PRE predictions using a set of TFBSs for proteins involved in the PcG/TrxG mediated regulation. D.mel, *D. melanogaster*; D.sim, *D. simulans*; D.sec, *D. sechellia*; D.yak, *D. yakuba*; D.ere, *D. erecta*; D.ana, *D. ananassae*; D.pse, *D. pseudoobscura* and D.per, *D. persimilis*.



Discusión

Las proteínas del grupo Polycomb están presentes en organismos de clados tan alejados filogenéticamente como protistas, plantas, hongos y animales, lo que implica que debieron originarse durante la aparición de los primeros eucariotas (Shaver et al. 2010), esto es, hace al menos 1.000 millones de años (Eme et al. 2014). La persistencia de las proteínas Polycomb desde hace tanto tiempo deja clara la importancia de este sistema de silenciamiento epigenético dentro de la maquinaria de regulación de la expresión génica en los eucariotas. No obstante, que el sistema se haya mantenido no implica que no haya sufrido cambios en cuanto al número de proteínas que lo integran (Boulay et al. 2011; Margueron y Reinberg 2011) o a los subprocesos específicos en los que participa (Payer y Lee 2008; Pu y Sung 2015). Es de esperar que estos cambios, a su vez, se vean respaldados por la selección adaptativa, actuando tanto sobre las secuencias de los genes Polycomb como sobre las secuencias reguladoras de sus genes diana.

Entre estos genes diana destacan especialmente el conjunto de genes Hox, claves en la formación de las distintas estructuras anatómicas durante el desarrollo embrionario. Cuando en un conjunto de organismos encontramos diferencias morfológicas debidas a la expresión diferencial de los genes Hox, cabe suponer que los genes Polycomb estén involucrados en el proceso. Sin duda, esto ha ocurrido a gran escala desde la radiación de los primeros eucariotas, pero también a escalas de tiempo menores, dentro del género *Drosophila* (Randsholt y Santamaria 2008), o incluso muy recientes, durante el proceso de especiación entre *D. subobscura* y *D. madeirensis* (Papacit, San Antonio y Prevosti 1991).

El contraste entre la conservación de este sistema altamente pleiotrópico y la evolución dinámica de sus componentes convierte la evolución molecular de los genes Polycomb en un interesante objeto de estudio. Ya existen trabajos sobre aspectos concretos de esta evolución, como la conservación de la estructura WD40 de la proteína ESC en diversos insectos (Ng et al. 1997), o el

mantenimiento de la función en el gen *Psc* pese a su elevada divergencia (Beh, Colwell y Francis 2012). No obstante, en la presente tesis doctoral se ha realizado por primera vez un estudio exhaustivo de la evolución molecular de los genes Polycomb en *Drosophila*, desde múltiples aproximaciones y metodologías. Los resultados, presentados a lo largo de cuatro artículos de investigación, incluyen: i) la descripción de su divergencia interespecífica y la confirmación de la hipótesis coevolutiva en respuesta a la actuación de sus proteínas formando complejos, ii) la identificación de un nuevo gen duplicado que ha sufrido un proceso de neofuncionalización, iii) la inferencia de selección adaptativa a distintas escalas temporales y iv) la caracterización de la región reguladora del gen *Hox Scr*.

Divergencia y coevolución de las proteínas Polycomb

Cuando un grupo de proteínas actúan conjuntamente en forma de complejos, los cambios aminoacídicos en una de ellas pueden afectar a su interacción con las demás y, por tanto, al rendimiento del complejo. En consecuencia, los dominios proteicos involucrados en la interacción entre proteínas se encuentran bajo una fuerte selección purificadora. Esto puede observarse al analizar la conservación aminoacídica de las proteínas Polycomb (Artículo 1, Figura 2). Los dominios funcionales previamente descritos en la bibliografía, en su mayor parte involucrados en interacciones entre proteínas, presentan un elevado grado de conservación en el alineamiento de las 15 especies de *Drosophila* estudiadas. Esta conservación diferencial de los dominios funcionales puede utilizarse a su vez para identificar nuevos dominios que no hayan sido previamente descritos. En el caso que nos ocupa, tanto el extremo N-terminal de la proteína SFMBT como la región comprendida entre los dominios NTD y C2H2 de la proteína SU(Z)12 presentan un alto grado de conservación aminoacídica, lo cual convierte a estas regiones en buenas candidatas a presentar algún tipo de función específica y deberían ser estudiadas experimentalmente.

Pese a que los dominios de interacción entre proteínas tienden a conservarse bajo la acción de la selección purificadora, puede ocurrir que se fijen mutaciones neutras por deriva genética o mutaciones beneficiosas sobre las que actúe la selección adaptativa. En este último caso, la selección también favorecerá la fijación de mutaciones compensatorias en las demás proteínas del complejo con tal de mantener una correcta interacción (Lovell y Robertson 2010). Se han desarrollado distintos métodos bioinformáticos para inferir este proceso de

coevolución molecular a partir de las correlaciones en las tasas evolutivas de las proteínas involucradas (Juan, Pazos y Valencia 2013). En la presente tesis hemos utilizado el método *ContextMirror* (Juan, Pazos y Valencia 2008) para generar un mapa coevolutivo del proteoma de las 12 especies secuenciadas y anotadas por el *Drosophila 12 Genomes Consortium* (Clark et al. 2007).

Los resultados muestran que, dentro de la red de correlaciones de todo el proteoma, 14 de las 17 proteínas Polycomb incluidas en el análisis se concentran en dos grupos perfectamente definidos (Artículo 1, Figura 5). Esto implica no solo que las correlaciones entre ellas superan el nivel de significación requerido, sino que además sus tasas evolutivas están más correlacionadas entre sí que con el resto de proteínas, evidenciando el proceso de coevolución entre los integrantes de los distintos complejos Polycomb. En el primero de los dos grupos se encuentran la proteína ASX junto con las integrantes del complejo Pcl-PRC2 a excepción de CAF1-55 y ESCL. En el segundo se agrupan la proteína CALYPSO, todas las integrantes de los complejos PRC1 y dRAF, y las integrantes del complejo PhoRC excepto PHOL. De las tres excepciones mencionadas, las correlaciones de PHOL no llegan a superar el nivel de significación requerido en el análisis y tanto CAF1-55 como ESCL se encuentran en la red coevolutiva, pero aisladas del resto de proteínas Polycomb.

Los resultados más inesperados son, por una parte, que las proteínas ASX y CALYPSO no presenten correlación directa pese a formar parte del mismo complejo PR-DUB (Scheuermann et al. 2010) y, por otra, que PHO y SFMBT, integrantes del complejo PhoRC, no coevolucionen con el complejo PRC2 pese a que las interacciones de PHO con E(Z) y ESC han sido experimentalmente probadas (Wang et al. 2004). Por lo demás, las buenas correlaciones obtenidas entre los complejos PhoRC y PRC1 refuerzan la hipótesis que defiende la importancia de PRC1 a la hora de reclutar al resto de complejos Polycomb y su capacidad para unirse a PhoRC en las secuencias PRE (Kahn et al. 2014). Finalmente, el análisis de la coevolución predice otras interacciones que no habían sido previamente descritas. Cabe destacar el caso de la proteína KDM2 que forma junto a SCE y PSC el complejo dRAF. A excepción del trabajo que detectó su co-inmunoprecipitación (Lagarou et al. 2008) no hay más información acerca de las interacciones entre KDM2 y otras proteínas Polycomb. Sin embargo, la coevolución detectada en este estudio indica que debe haber una interacción directa no solo entre KDM2 y las otras dos integrantes de dRAF, sino también con proteínas de los complejos PhoRC y PRC1.

Aunque los resultados evidencian que las proteínas Polycomb están más correlacionadas entre sí que con el resto del proteoma de *Drosophila*, los dos grupos identificados no están aislados sino que forman parte de una inmensa red de correlaciones. Por tanto, hay un buen número de proteínas que parecen interactuar con el grupo Polycomb pese a no tener una aparente relación funcional. La excepción la constituye la proteína Hox SCR, conocida diana de la regulación mediada por el grupo Polycomb (Gindhart y Kaufman 1995). Los resultados muestran que SCR parece coevolucionar con integrantes de los complejos PhoRC, PRC1, dRAF y PR-DUB, lo cual sugiere cierta interacción tal vez involucrada en su propia autorregulación.

Identificación de un nuevo parálogo: *Caf1-55dup*

No es sorprendente que sean las proteínas CAF1-55, PHOL y ESCL las únicas que no muestran indicios de coevolucionar con el resto de integrantes de los complejos Polycomb. *Caf1-55* es con diferencia el gen Polycomb más conservado en la filogenia de *Drosophila* (Artículo 2, Figura 1) y la proteína codificada participa, además de en PRC2, en otros complejos remodeladores de la cromatina (Song, Garlick y Kingston 2008). Por su parte, los genes *Pho* y *Esc* son parálogos de *Pho* y *Esc*, respectivamente, y las proteínas que codifican no forman parte de los complejos principales sino que actúan como alternativa a PHO y ESC en determinados tejidos o fases del desarrollo (Wang et al. 2004; Ohno et al. 2008). Por tanto, en estos genes surgidos por duplicación es de esperar que las presiones selectivas que guían su coevolución con el resto del grupo Polycomb sean más leves. Un caso similar sería el del gen *Ph-d*, parálogo de *Ph-p*, aunque no se incluye en el estudio de coevolución dado que está presente en solo 8 de las 12 especies de *Drosophila* anotadas en FlyBase (Artículo 2, Figura 4). En base a los árboles filogenéticos de estas tres parejas de genes duplicados (Artículo 2, Figura S5) y a la información disponible sobre sus diferentes funciones y niveles de expresión, puede deducirse que sufrieron un proceso de neofuncionalización. Tras la presente tesis doctoral, a estas tres duplicaciones dentro del grupo Polycomb debe añadirse una última y más reciente: precisamente la del gen *Caf1-55*, cuyo parálogo hemos denominado *Caf1-55dup*.

Los resultados muestran que *Caf1-55dup* se encuentra únicamente en las especies del grupo *obscura*, originándose en su ancestro común hace aproximadamente 18 millones de años, en un evento de duplicación mediado

por ADN que afectó únicamente a la región génica de *Caf1-55*. La secuenciación de su ARN mensajero confirma que el gen se expresa (Artículo 2, Figura S2) y los datos de ARN-seq para *D. pseudoobscura* disponibles en FlyBase indican que *Caf1-55* y *Caf1-55dup* difieren en sus niveles y patrones de expresión (Artículo 2, Figura S6). Mientras el primero aparece ubicuamente en todos los tejidos, la expresión del segundo es mucho más leve y restringida a testículos y ovarios. Esto probablemente es debido a que el evento de duplicación no incluyó la región reguladora de *Caf1-55*, por lo que *Caf1-55dup* debió adoptar un promotor preexistente en el sitio de inserción o incluso aguardar a su evolución *de novo*. El hecho de que desde su origen *Caf1-55dup* tuviera un patrón de expresión distinto o no se expresara en absoluto explicaría que la selección pudiera discriminar entre ambas copias inmediatamente y que, por tanto, no encontremos el menor atisbo de relajación de la selección purificadora en *Caf1-55*.

A diferencia de *Caf1-55*, *Caf1-55dup* presenta unos niveles de divergencia bastante elevados incluso entre especies próximas. No obstante, el análisis detallado de su secuencia aminoacídica revela que mantiene en todas las especies la estructura β -*propeller* con 7 repeticiones WD40, presente en *Caf1-55* y característica de las proteínas de la familia WD40 (Wang et al. 2015). De hecho, la inmensa mayoría de aminoácidos involucrados en esta estructura se hallan conservados, mientras que aquellos residuos involucrados en la interacción con otras proteínas presentan muchos más cambios respecto a *Caf1-55* y entre los propios ortólogos de *Caf1-55dup* (Artículo 2, Figura3). Dado que la familia de proteínas WD40 es una de las más extensas en los eucariotas (Stirnemann et al. 2010) y que, de hecho, CAF1-55 forma parte de diversos complejos además de PRC2, la conservación de la estructura en CAF1-55DUP no nos permite afirmar que este nuevo duplicado sea un miembro del grupo Polycomb. Por otro lado, la conservación diferencial de los residuos estructurales respecto a los de interacción con otras proteínas implica, en primer lugar, que el gen duplicado probablemente es funcional (a falta de la detección de la proteína), ya que si se tratara de un pseudogén la selección purificadora no actuaría sobre los aminoácidos involucrados en su estructura. En segundo lugar, nos indica que *Caf1-55dup*, como el resto de parálogos Polycomb, seguramente ha evolucionado bajo un proceso de neofuncionalización. Como veremos a continuación, esta hipótesis a su vez está sustentada por la detección de selección adaptativa en este nuevo duplicado.

Inferencia de selección adaptativa

Los 17 genes Polycomb incluidos en el estudio de coevolución, junto a *Caf1-55dup*, han sido analizados mediante metodologías de máxima verosimilitud con el fin de detectar tanto aceleraciones en las tasas evolutivas de linajes concretos como episodios puntuales de selección adaptativa (Yang 2007; Smith et al. 2015). En este análisis, los alineamientos de partida incluyen las secuencias de 15 especies de *Drosophila*, las 12 anotadas y disponibles en FlyBase más las tres especies del grupo *obscura*, *D. subobscura*, *D. madeirensis* y *D. guanche*, secuenciadas en este trabajo. En el estudio de la duplicación de *Caf1-55* el alineamiento llega a las 16 especies al haber incluido *D. miranda*. Los resultados de aplicar los *Branch models* (Introducción, Figura 3) a cada uno de los genes Polycomb muestran que siete de ellos comparten un patrón caracterizado por una aceleración de la tasa de cambio aminoacídica en la rama ancestral del grupo *obscura*, lo que provoca que la ω (d_n/d_s) en esta rama sea significativamente más elevada que en el resto de ramas de la filogenia (Artículo 1, Figura 4).

Los siete genes que presentan esta aceleración significativa en el linaje ancestral del grupo *obscura*, por orden de mayor a menor ω , son: *Esc*, *Sfmbt*, *Pcl*, *E(z)*, *Su(z)12*, *Psc* y *Kdm2*. De hecho, en los cuatro primeros puede observarse claramente dicha aceleración en las reconstrucciones filogenéticas correspondientes (Artículo 1, Figura 3). Curiosamente, junto al extremadamente conservado *Caf1-55*, cuatro de estos siete genes conforman el complejo Pcl-PRC2, lo que sugiere que el aumento de cambios aminoacídicos en la rama ancestral del grupo *obscura* podría contribuir al proceso de coevolución detectado. Si tenemos en cuenta que *Caf1-55dup* se originó también en el ancestro del grupo *obscura*, así como su elevada divergencia aminoacídica tras la duplicación (Artículo 2, Figura 2), cabe contemplar la posibilidad de que el proceso de neofuncionalización esté a su vez ligado a la coevolución de los integrantes del complejo Pcl-PRC2. Esto implicaría que la proteína CAF1-55DUP no solo sigue formando parte de la familia WD40 sino que además mantiene una función relacionada con el grupo Polycomb.

En cuanto a la detección de selección adaptativa en ramas específicas de la filogenia, los resultados de aplicar los *Branch-site models* (Introducción, Figura 3) han desvelado episodios puntuales de selección en el linaje ancestral del grupo *obscura* de los genes *Sfmbt*, *E(z)* y *Caf1-55dup*. Posteriormente, mediante el método BEB (*Bayes Empirical Bayes*) (Yang, Wong y Nielsen 2005) se han identificado aquellos codones concretos con una probabilidad significativa

de presentar una $\omega > 1$ (Artículo 1, Figura 2). Los seis codones bajo selección adaptativa identificados en $E(z)$ están localizados en sus dominios de interacción con otras proteínas Polycomb, lo que refuerza la idea de que su coevolución estuvo guiada por dicha selección. Respecto a *Sfmbt*, cuatro de los siete codones bajo selección adaptativa se ubican en su extremo N-terminal, precisamente la misma región que ha sido identificada, en base a su conservación, como una buena candidata a constituir un nuevo dominio funcional. El hecho de que tanto $E(z)$ como *Sfmbt* presenten una aceleración de la tasa de cambio aminoacídica en la rama ancestral del grupo *obscura*, guiada por la acción de la selección adaptativa, podría interpretarse como un indicio de coevolución entre ambos genes. No obstante, el mapa coevolutivo no refleja esta interacción ni ninguna otra entre los complejos Pcl-PRC2 y PhoRC. Por tanto, o bien su coevolución es demasiado débil como para que sea detectada por el método *ContextMirror*, o la evolución de ambos genes es independiente pese a mostrar un patrón similar. Finalmente, la identificación de 22 codones bajo selección adaptativa en el gen duplicado *Caf1-55dup* (Artículo 2, Figura S3), junto a la extrema selección purificadora que ha actuado sobre *Caf1-55* y el hecho de que no hayamos detectado indicios de relajación selectiva en ninguno de los dos genes, no hacen sino confirmar el proceso de neofuncionalización mencionado anteriormente.

La búsqueda de selección adaptativa en los genes del grupo Polycomb no se ha limitado al estudio de divergencia interespecífica discutido hasta ahora, sino que también se ha llevado a cabo a una escala más reciente, poniendo el foco en la variación intraespecífica de *D. subobscura*. Los genes estudiados son los mismos que en el estudio de divergencia a excepción de los dos integrantes de PR-DUB e incluyendo en esta ocasión a *Ph-d*, parálogo de *Ph-p*. Partiendo de los alineamientos de las regiones génicas en un promedio de 16 líneas altamente consanguíneas, se buscaron desviaciones del modelo neutro estándar en los patrones de polimorfismo nucleotídico. Con tal de discriminar entre las desviaciones provocadas por la selección adaptativa y aquellas causadas por los cambios demográficos de la especie, el nivel de significación de los estimadores y test de neutralismo utilizados se ha calculado en base a simulaciones ajustadas al modelo de expansión poblacional establecido previamente para *D. subobscura* (Pratdesaba, Segarra y Aguadé 2015). Los resultados muestran que la totalidad de los genes estudiados se ajustan al equilibrio mutación-deriva teniendo en cuenta la expansión poblacional. Únicamente cuando comparamos, mediante el test MK (McDonald y Kreitman 1991), las posiciones sinónimas y no sinónimas dentro de *D. subobscura* y respecto a la especie próxima *D. guanche* hallamos indicios de selección adaptativa en dos genes: *Su(z)12* y, de nuevo, *Caf1-55dup*.

Respecto a los niveles de variación, la mayoría de los genes estudiados presentan una proporción entre polimorfismo sinónimo y no sinónimo (π_a/π_s) que se ajusta a la ratio equivalente observada en el estudio de divergencia (ω). Así, el gen *Caf1-55* es el que sufre una selección purificadora más intensa tanto a nivel intraespecífico como interespecífico. Por su parte, genes como *Pcl*, *Psc* o *Ph-p* presentan las mayores ratios π_a/π_s en *D. subobscura* y unos elevados valores de ω a lo largo de la filogenia de *Drosophila*, de los que se infiere una menor selección purificadora y que, a su vez, concuerda con su posible condición de genes estructuralmente desordenados, confirmada analíticamente en *Psc*, *Ph-p* y *Asx*. En estos genes la conservación de su función no recae en la integridad de su secuencia aminoacídica sino en el mantenimiento de regiones con una determinada carga eléctrica (Beh, Colwell y Francis 2012).

El gen *Pho* presenta solo cuatro polimorfismos no sinónimos y una total ausencia de variación sinónima, por lo que constituye un caso especial entre los genes Polycomb. Esto se observa claramente en el resultado del test HKA (Hudson, Kreitman y Aguadé 1987), el cual muestra en *Pho* un déficit significativo de polimorfismo respecto a su divergencia con *D. guanche* (Artículo 3, Figura S1). Este déficit de variación intraespecífica se debe probablemente a la localización del gen *Pho* en el cromosoma puntiforme de *D. subobscura* (Artículo 3, Figura 1). Se ha observado previamente que el cromosoma puntiforme de *D. melanogaster* presenta niveles muy bajos de recombinación, provocando que cualquier mutación bajo selección positiva o negativa comporte un barrido selectivo que elimina la variabilidad en amplias regiones del cromosoma (Wang et al. 2002). El déficit de polimorfismo observado en el gen *Pho*, junto con la ausencia de eventos de recombinación en toda su región génica, indican por primera vez que el fenómeno observado en *D. melanogaster* puede darse también en *D. subobscura*.

La evolución del gen *Pho* no solo presenta particularidades a nivel intraespecífico sino también en cuanto a su divergencia. Parte de su secuencia nucleotídica es tan variable entre las distintas especies del género *Drosophila* que constituye el único gen Polycomb estudiado en esta tesis para el que la homología de varios de sus exones no ha podido ser totalmente confirmada (Artículo 1, Figura 1). Sin embargo, su principal dominio funcional, compuesto por 4 dedos de zinc C2H2, está extremadamente conservado en toda la filogenia. Este dominio, compartido con su parálogo *Phol*, es el que le confiere la capacidad de unión al ADN en las secuencias reguladoras PRE presentes en los genes regulados por las proteínas Polycomb. Como ocurre con el gen *Trl*, también implicado en la unión

a las secuencias PRE (Cléard et al. 2016), la conservación del dominio de unión a ADN podría implicar que el sitio de unión reconocido por éste también se ha mantenido conservado en las especies de *Drosophila* estudiadas.

Caracterización de la región reguladora del gen Hox *Scr*

Como hemos visto en el análisis con *ContextMirror*, la proteína SCR parece coevolucionar con un buen número de proteínas Polycomb. Teniendo en cuenta que el gen *Scr* es una de sus dianas, cabe esperar que la evolución de su región reguladora también se vea influenciada por la evolución de los propios genes Polycomb. Para estudiar esta posibilidad se han analizado, mediante el programa MOTEVO (Arnold et al. 2012), las ~70 kb que constituyen la región reguladora de *Scr* en busca de módulos reguladores en *cis* (CRMs, por sus siglas en inglés), centrándonos no solamente en los elementos PRE, sino también en los llamados *Initiator Elements* (IEs), responsables del reclutamiento de factores de segmentación. MOTEVO basa la predicción de CRMs en su conservación evolutiva, lo que nos permite observar si los elementos reguladores identificados se mantienen a lo largo de las ocho especies de los grupos *melanogaster* y *obscura* que se han incluido en el alineamiento de partida. Para respaldar los resultados se han recopilado todos los datos experimentales de *chromatin immunoprecipitation* (ChIP) disponibles en *D. melanogaster*, tanto de unión de proteínas Polycomb y sus respectivos cofactores, como de factores de segmentación y otras proteínas involucradas en la formación de *Insulators* o en la remodelación de la cromatina (Artículo 4, Figura 1).

Los resultados de combinar la recopilación de datos ChIP con el análisis bioinformático consiguen, no solo identificar la inmensa mayoría de los CRMs de *Scr* descubiertos hasta la fecha y algún elemento nuevo, sino acotar mejor su localización, así como redefinir algunas de sus características (Artículo 4, Figura 6). Entre los elementos identificados por ambas aproximaciones destacan IE4/PRE5, caracterizado por reclutar tanto proteínas Polycomb como factores de segmentación; la pareja PRE4 y PRE9, ubicados en los dos inicios de la transcripción alternativos de *Scr* y que dividen lo que anteriormente se consideraba un único CRM (2.9BR); y los elementos IE3 y PRE2 que, como en el caso de PRE4 y PRE9, acotan y dividen en dos elementos completamente distintos lo que antes se consideraba un solo CRM de 10 kb (10.0XX).

MOTEVO también se ha utilizado para comprobar si los llamados *tethering elements* identificados previamente en *D. melanogaster* (Calhoun y Levine 2003) se mantienen en las demás especies estudiadas. Estos dos elementos, compuestos por entre cuatro y ocho repeticiones palindrómicas, parecen estar implicados en el plegamiento de la región reguladora con el fin de acercar los CRMs distales a la región promotora de *Scr*. Los resultados no solo confirman que ambos mantienen al menos cuatro repeticiones en todas las especies sino que muestran que, de hecho, forman parte de los elementos PRE4 y PRE2 identificados en este trabajo. En vista de este resultado cabía plantearse si la presencia de *tethering elements* era una característica común a otros elementos PRE en *Drosophila*. Sin embargo, la búsqueda de estas repeticiones en los elementos PRE presentes a lo largo de las regiones genómicas correspondientes a los complejos Bithorax y Antennapedia no ha dado resultados positivos, por lo que deducimos que la combinación de PREs y *tethering elements* detectada en la región reguladora de *Scr* no puede extrapolarse de momento a otras regiones reguladoras.

Todos los elementos reguladores identificados en este trabajo parecen estar conservados en las ocho especies analizadas con MOTEVO a excepción de PRE9, ausente en *D. ananassae*, *D. pseudoobscura* y *D. persimilis* (Artículo 4, Figura S2). Esto parece indicar que la región reguladora de *Scr* no ha sufrido grandes cambios en los ~24 millones de años que separan el grupo *melanogaster* del grupo *obscura* (Obbard et al. 2012), al menos en cuanto al número y ubicación de los elementos reguladores que lo conforman. No obstante, no puede descartarse que se hayan producido variaciones en la composición de cada uno de estos elementos reguladores. Respecto a PRE9, el hecho de que esté ausente en *D. ananassae* además de en las dos especies del grupo *obscura* parece indicar que no se trata de una pérdida en estas especies sino una ganancia en las del subgrupo *melanogaster*.

Conclusions

- 1 El grado de conservación en el género *Drosophila* de los 17 genes Polycomb estudiados es muy variable. La fuerte conservación del gen *Caf1-55* contrasta con la elevada divergencia de genes como *Psc* y *Ph-p*, caracterizados por ser estructuralmente desordenados.
- 2 El análisis de la divergencia a lo largo de cada una de las proteínas Polycomb muestra una buena relación entre regiones conservadas y dominios funcionales, aunque también se han identificado regiones con baja divergencia que no se corresponden con ningún dominio previamente descrito.
- 3 Las proteínas Polycomb muestran indicios claros de coevolución en el género *Drosophila*, tal como podía predecirse para unas proteínas que actúan formando complejos multiproteicos.
- 4 La red de correlaciones evolutivas inferida en el análisis de la coevolución mediante el método *ContextMirror* confirma interacciones que ya habían sido caracterizadas experimentalmente y predice nuevas interacciones, tanto entre proteínas de un mismo complejo como entre proteínas de complejos distintos.
- 5 Siete de los genes Polycomb estudiados, de los que cuatro codifican proteínas del complejo Pcl-PRC2, presentan un incremento significativo de la tasa de substitución aminoacídica en el linaje ancestral de las especies del grupo *obscura*.

- 6 En los genes *Sfmbt* y *E(z)* se ha podido confirmar que dicho incremento se debe a episodios puntuales de selección adaptativa.
- 7 Cuatro de los siete codones con una elevada probabilidad de haber evolucionado bajo selección adaptativa en el gen *Sfmbt* se agrupan en el inicio de la región codificadora, caracterizada por presentar una gran conservación en las posiciones no sinónimas, lo que sugiere que esta región codifica un nuevo dominio funcional.
- 8 El gen *Caf1-55* está duplicado en las especies del grupo *obscura* de *Drosophila*. Se ha inferido que la duplicación se produjo hace ~18 millones de años y que, tras su origen, la nueva copia (*Caf1-55dup*) sufrió un proceso de neofuncionalización guiado por selección adaptativa, manteniendo su estructura WD40.
- 9 En el gen *Caf1-55dup* también se ha detectado la acción de selección adaptativa en una escala temporal mucho más reciente, durante la divergencia de *D. subobscura* y *D. guanche*.
- 10 Los niveles de polimorfismo nucleotídico de los genes Polycomb en la especie *D. subobscura* son muy variables, siendo especialmente destacable la variación en la ratio entre polimorfismo no sinónimo y sinónimo que oscila entre 0 en *Caf1-55* y 0.1704 en *Pcl*.
- 11 El gen *Pho*, localizado en el cromosoma puntiforme, presenta un déficit significativo de polimorfismo, así como una baja diversidad haplotípica y un desequilibrio de ligamiento extremo. Este nivel y patrón de polimorfismo indica que este cromosoma de *D. subobscura* también presenta la fuerte reducción de la recombinación descrita en el cromosoma homólogo de *D. melanogaster*.
- 12 El gen *Su(z)12* muestra un desajuste en la ratio de polimorfismo y divergencia entre las posiciones sinónimas y no sinónimas, lo que indica el efecto de la selección adaptativa sobre este gen durante la divergencia de *D. subobscura* y *D. guanche*.

- 13 La estrategia basada en combinar la predicción bioinformática de CRMs con datos experimentales de ChIP ha permitido mejorar la caracterización de los elementos PRE e IE de la región reguladora del gen *Scr*. Varios elementos reguladores previamente descritos han sido redefinidos, acotados e incluso subdivididos en elementos independientes.
- 14 Los dos *tethering elements* de la región reguladora *Scr* se superponen con sendos elementos PRE identificados en este trabajo. No obstante, esta asociación descrita aquí por primera vez no puede extrapolarse a otras regiones reguladoras en los complejos Antennapedia y Bithorax.
- 15 La gran mayoría de los elementos reguladores identificados están conservados a lo largo de las ocho especies de los grupos *melanogaster* y *obscura* incluidas en el análisis. La única excepción es PRE9 que parece haberse originado en el ancestro del subgrupo *melanogaster*.

Bibliografia

Bibliografía

- Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD et al. (2000) The genome sequence of *Drosophila melanogaster*. *Science* 287: 2185–2195.
- Alvarez-Ponce D, Guirao-Rico S, Orengo DJ, Segarra C, Rozas J et al. (2012) Molecular population genetics of the insulin/TOR signal transduction pathway: a network-level analysis in *Drosophila melanogaster*. *Mol. Biol. Evol.* 29: 123–132.
- Arnold P, Erb I, Pachkov M, Molina N, Van Nimwegen E (2012) MotEvo: integrated bayesian probabilistic methods for inferring regulatory sites and motifs on multiple alignments of DNA sequences. *Bioinformatics* 28: 487–494.
- Assis R, Bachtrög D (2013) Neofunctionalization of young duplicate genes in *Drosophila*. *Proc. Natl. Acad. Sci. U. S. A.* 110: 17409–17414.
- Auton A, Abecasis GR, Altshuler DM, Durbin RM, Bentley DR et al. (2015) A global reference for human genetic variation. *Nature* 526: 68–74.
- Avery OT, Macleod CM, McCarty M (1944) Studies on the chemical nature of the substance inducing transformation of pneumococcal types: induction of transformation by a desoxyribonucleic acid fraction isolated from pneumococcus type III. *J. Exp. Med.* 79: 137–158.
- Bantignies F, Cavalli G (2011) Polycomb group proteins: repression in 3D. *Trends Genet.* 27: 454–464.
- Beckenbach AT, Prevosti A (1986) Colonization of North America by the european species, *Drosophila subobscura* and *D. ambigua*. *Am. Midl. Nat.* 115: 10–18.
- Beh LY, Colwell LJ, Francis NJ (2012) A core subunit of Polycomb repressive complex 1 is broadly conserved in function but not primary sequence. *Proc. Natl. Acad. Sci. U. S. A.* 109: E1063–E1071.
- Beisswanger S, Stephan W (2008) Evidence that strong positive selection drives neofunctionalization in the tandemly duplicated polyhomeotic genes in *Drosophila*. *Proc. Natl. Acad. Sci. U. S. A.* 105: 5447–5452.
- Bender W, Fitzgerald DP (2002) Transcription activates repressed domains in the *Drosophila* bithorax complex. *Development* 129: 4923–4930.
- Bennett JH, Fisher R (1930) The Genetical Theory of Natural Selection. Oxford: Clarendon Press
- Boulay G, Rosnoblet C, Guérardel C, Angrand P-O, Leprince D (2011) Functional characterization of human Polycomb-like 3 isoforms identifies them as components of distinct EZH2 protein complexes. *Biochem. J.* 434: 333–342.
- Brneie D, Prevosti A, Budnik M, Monclus M, Ocana J (1981) Colonization of *Drosophila subobscura* in Chile. I. First population and cytogenetic studies. *Genetica* 56: 3–9.
- Brown JL, Fritsch C, Mueller J, Kassiss JA (2003) The *Drosophila pho-like* gene encodes a YY1-related DNA binding protein that is redundant with pleiohomeotic in homeotic gene silencing. *Development* 130: 285–294.
- Brown JL, Kassiss JA (2013) Architectural and functional diversity of Polycomb group response elements in *Drosophila*. *Genetics* 195: 407–419.
- Calhoun VC, Levine M (2003) Long-range enhancer-promoter interactions in the *Scr-Antp* interval of the *Drosophila* Antennapedia complex. *Proc. Natl. Acad. Sci. U. S. A.* 100: 9878–9883.

- Chanvivattana Y, Bishopp A, Schubert D, Stock C, Moon Y-H et al. (2004) Interaction of Polycomb-group proteins controlling flowering in *Arabidopsis*. *Development* 131: 5263–5276.
- Chen S, Birve A, Rasmuson-Lestander Å (2008) In vivo analysis of *Drosophila* SU(Z)12 function. *Mol. Genet. Genomics* 279: 159–170.
- Clark AG, Eisen MB, Smith DR, Bergman CM, Oliver B et al. (2007) Evolution of genes and genomes on the *Drosophila* phylogeny. *Nature* 450: 203–218.
- Cléard F, Wolle D, Taverner AM, Aoki T, Deshpande G et al. (2016) Different evolutionary strategies to conserve chromatin boundary function in the Bithorax complex. *Genetics* 205: 589–603.
- Cunningham MD, Brown JL, Kassiss JA (2010) Characterization of the Polycomb group response elements of the *Drosophila melanogaster* *invected* locus. *Mol. Cell. Biol.* 30: 820–828.
- Czermin B, Melfi R, McCabe D, Seitz V, Imhof A et al. (2002) *Drosophila* Enhancer of Zeste/ESC complexes have a histone H3 methyltransferase activity that marks chromosomal Polycomb sites. *Cell* 111: 185–196.
- Darwin C (1859) *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life*. London: J. Murray
- Denoeud F, Carretero-Paulet L, Dereeper A, Droc G, Guyot R et al. (2014) The coffee genome provides insight into the convergent evolution of caffeine biosynthesis. *Science* 345: 1181–1184.
- Dietrich N, Bracken AP, Trinh E, Schjerling CK, Koseki H et al. (2007) Bypass of senescence by the Polycomb group protein CBX8 through direct binding to the INK4A-ARF locus. *EMBO J.* 26: 1637–1648.
- Dobzhansky T (1937) *Genetics and the Origin of Species*. New York: Columbia Univ. Press
- Eme L, Sharpe SC, Brown MW, Roger AJ (2014) On the age of eukaryotes: evaluating evidence from fossils and molecular clocks. *Cold Spring Harb. Perspect. Biol.* 6: a016139.
- Fay JC, Wu CI (2000) Hitchhiking under positive Darwinian selection. *Genetics* 155: 1405–1413.
- Fischle W, Wang Y, Jacobs SA, Kim Y, Allis CD et al. (2003) Molecular basis for the discrimination of repressive methyl-lysine marks in histone H3 by Polycomb and HP1 chromodomains. *Genes Dev.* 17: 1870–1881.
- 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.
- Franklin RE, Gosling RG (1953) Molecular configuration in sodium thymonucleate. *Nature* 171: 740–741.
- Fritsch C, Brown JL, Kassiss JA, Müller J (1999) The DNA-binding Polycomb group protein pleiohomeotic mediates silencing of a *Drosophila* homeotic gene. *Development* 126: 3905–3913.
- Fu YX (1997) Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. *Genetics* 147: 915–925.

Bibliografía

- Fu YX, Li WH (1993) Statistical tests of neutrality of mutations. *Genetics* 133: 693–709.
- Gaytán de Ayala Alonso A, Gutiérrez L, Fritsch C, Papp B, Beuchle D et al. (2007) A genetic screen identifies novel Polycomb group genes in *Drosophila*. *Genetics* 176: 2099–2108.
- Gindhart JG, Kaufman TC (1995) Identification of Polycomb and trithorax group responsive elements in the regulatory region of the *Drosophila* homeotic gene *Sex combs reduced*. *Genetics* 139: 797–814.
- Grimm C, Matos R, Ly-Hartig N, Steuerwald U, Lindner D et al. (2009) Molecular recognition of histone lysine methylation by the Polycomb group repressor *dSfmbt*. *EMBO J.* 28: 1965–1977.
- Guirao-Rico S, Aguadé M (2011) Molecular evolution of the ligands of the insulin-signaling pathway: dilp genes in the genus *Drosophila*. *Mol. Biol. Evol.* 28: 1557–1560.
- Gutierrez L, Oktaba K, Scheuermann JC, Gambetta MC, Ly-Hartig N et al. (2012) The role of the histone H2A ubiquitinase Sce in Polycomb repression. *Development* 139: 117–127.
- Hakes L, Lovell SC, Oliver SG, Robertson DL (2007) Specificity in protein interactions and its relationship with sequence diversity and coevolution. *Proc. Natl. Acad. Sci. U. S. A.* 104: 7999–8004.
- Haldane JBS (1990) *The Causes of Evolution*. London: Princeton University Press
- Hales KG, Korey CA, Larracuente AM, Roberts DM, Adams MD et al. (2015) Genetics on the fly: a primer on the *Drosophila* model system. *Genetics* 201: 815–842.
- Hanfstringl U, Berry A, Kellogg EA, Costa JT, Rüdiger W et al. (1994) Haplotypic divergence coupled with lack of diversity at the *Arabidopsis thaliana* alcohol dehydrogenase locus: roles for both balancing and directional selection? *Genetics* 138: 811–828.
- Hauenschild A, Ringrose L, Altmutter C, Paro R, Rehmsmeier M (2008) Evolutionary plasticity of Polycomb/Trithorax response elements in *Drosophila* species. *PLoS Biol.* 6: e261.
- Hershey AD, Chase M (1952) Independent functions of viral protein and nucleic acid in growth of bacteriophage. *J. Gen. Physiol.* 36: 39–56.
- Herzog VA, Lempradl A, Trupke J, Okulski H, Altmutter C et al. (2014) A strand-specific switch in noncoding transcription switches the function of a Polycomb/Trithorax response element. *Nat. Genet.* 46: 973–981.
- Hudson RR, Kreitman M, Aguadé M (1987) A test of neutral molecular evolution based on nucleotide data. *Genetics* 116: 153–159.
- Hughes CL, Kaufman TC (2002) Hox genes and the evolution of the arthropod body plan. *Evol. Dev.* 4: 459–499.
- Hur M-W, Laney JD, Jeon SH, Ali J, Biggin MD (2002) Zeste maintains repression of Ubx transgenes: Support for a new model of Polycomb repression. *Development* 129: 1339–1343.
- Huxley J (1942) *Evolution: The Modern Synthesis*. London: George Allen & Unwin
- Juan D, Pazos F, Valencia A (2008) High-confidence prediction of global interactomes based on genome-wide coevolutionary networks. *Proc. Natl. Acad. Sci. U. S. A.* 105: 934–939.

- Juan D, Pazos F, Valencia A (2013) Emerging methods in protein co-evolution. *Nat. Rev. Genet.* 14: 249–261.
- Kahn TG, Dorafshan E, Schultheis D, Zare A, Stenberg P et al. (2016) Interdependence of PRC1 and PRC2 for recruitment to Polycomb Response Elements. *Nucleic Acids Res.* 3: 10132–10149.
- Kahn TG, Stenberg P, Pirrotta V, Schwartz YB (2014) Combinatorial interactions are required for the efficient recruitment of Pho repressive complex (PhoRC) to Polycomb response elements. *PLoS Genet.* 10: e1004495.
- Kassis JA, Brown JL (2013) Polycomb group response elements in *Drosophila* and vertebrates. *Adv. Genet.* 81: 83–118.
- Kavi HH, Birchler JA (2009) *Drosophila* KDM2 is a H3K4me3 demethylase regulating nucleolar organization. *BMC Res. Notes* 2: e217.
- Kimura M (1968) Evolutionary rate at the molecular level. *Nature* 217: 624–626.
- Klymenko T, Papp B, Fischle W, Kocher T, Schelder M et al. (2006) A Polycomb group protein complex with sequence-specific DNA-binding and selective methyl-lysine-binding activities. *Genes Dev.* 20: 1110–1122.
- Koppens M, Van Lohuizen M (2015) Context-dependent actions of Polycomb repressors in cancer. *Oncogene* 35: 1341–1352
- Kreitman M (1983) Nucleotide polymorphism at the *alcohol dehydrogenase* locus of *Drosophila melanogaster*. *Nature* 304: 412–417.
- Kreitman M, Hudson RR (1991) Inferring the evolutionary histories of the *Adh* and *Adh-dup* loci in *Drosophila melanogaster* from patterns of polymorphism and divergence. *Genetics* 127: 565–582.
- Krimbas CB (1993) *Drosophila subobscura*: Biology, Genetics and Inversion Polymorphism. Hamburg: Verlag Dr. Kovac
- Krimbas CB, Powell JR (1992) *Drosophila* Inversion Polymorphism. Boca Raton: CRC Press
- Kurzchals RL, Tie F, Stratton CA, Harte PJ (2008) *Drosophila* ESC-like can substitute for ESC and becomes required for Polycomb silencing if ESC is absent. *Dev. Biol.* 313: 293–306.
- Lagarou A, Mohd-Sarip A, Moshkin YM, Chalkley GE, Bezstarosti K et al. (2008) dKDM2 couples histone H2A ubiquitylation to histone H3 demethylation during Polycomb group silencing. *Genes Dev.* 22: 2799–2810.
- Lanzuolo C, Orlando V (2012) Memories from the Polycomb group proteins. *Annu. Rev. Genet.* 46: 561–589.
- Lazzaro BP (2008) Natural selection on the *Drosophila* antimicrobial immune system. *Curr. Opin. Microbiol.* 11: 284–289.
- Lee H-G, Kahn TG, Simcox A, Schwartz YB, Pirrotta V (2015) Genome-wide activities of Polycomb complexes control pervasive transcription. *Genome Res.* 25: 1170–1181.
- Lo SM, Ahuja NK, Francis NJ (2009) Polycomb group protein Suppressor 2 of zeste is a functional homolog of Posterior Sex Combs. *Mol. Cell. Biol.* 29: 515–525.

Bibliografía

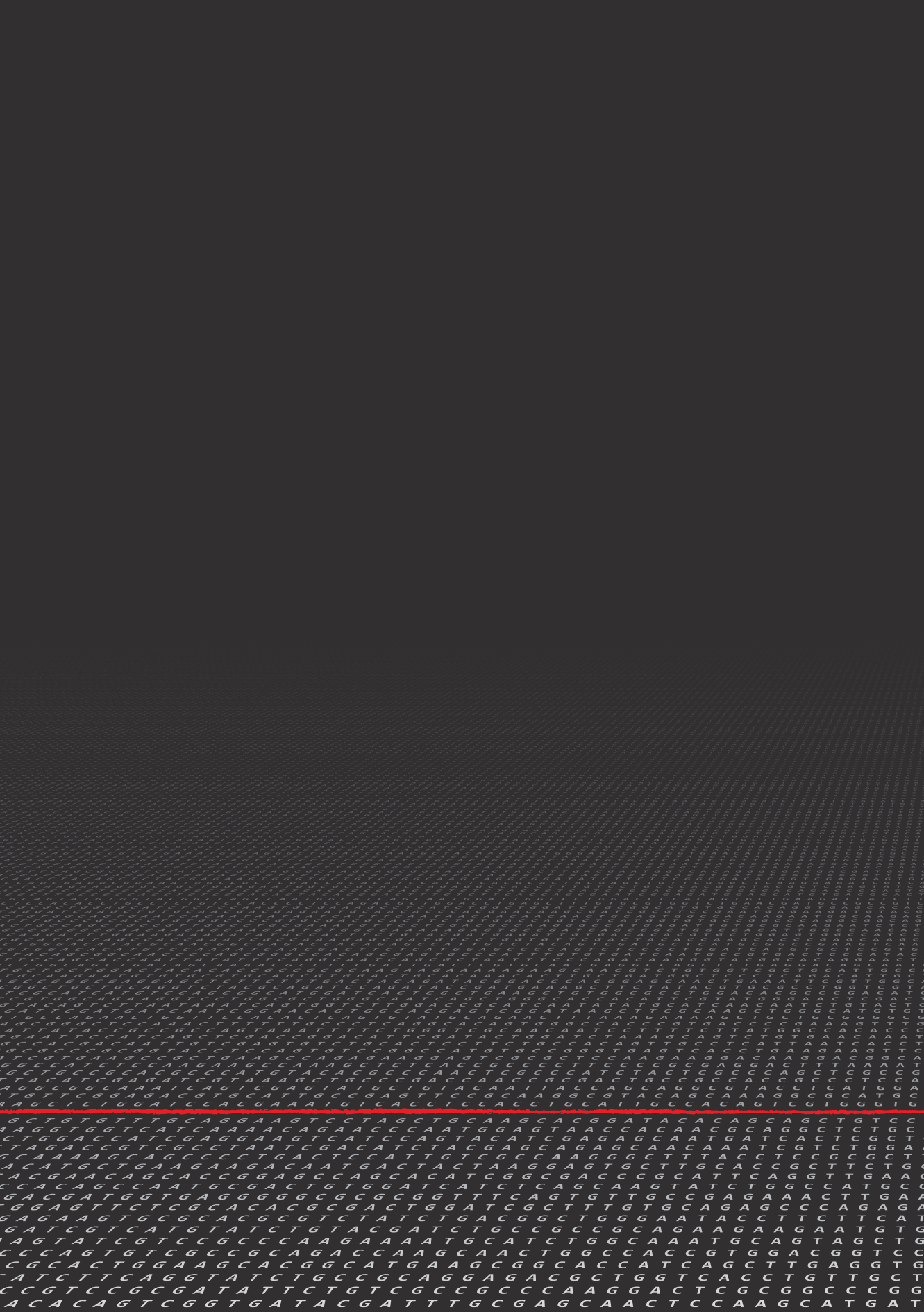
- Lo SM, McElroy KA, Francis NJ (2012) Chromatin modification by PSC occurs at one PSC per nucleosome and does not require the acidic patch of histone H2A. *PLoS One* 7: e47162.
- Lovell SC, Robertson DL (2010) An integrated view of molecular coevolution in protein-protein interactions. *Mol. Biol. Evol.* 27: 2567–2575.
- Mackay TFC, Richards S, Stone EA, Barbadilla A, Ayroles JF et al. (2012) The *Drosophila melanogaster* genetic reference panel. *Nature* 482: 173–178.
- Mahmoudi T, Zuijderduijn LMP, Mohd-Sarip A, Verrijzer CP (2003) GAGA facilitates binding of Pleiohomeotic to a chromatinized Polycomb response element. *Nucleic Acids Res.* 31: 4147–4156.
- Margueron R, Reinberg D (2011) The Polycomb complex PRC2 and its mark in life. *Nature* 469: 343–349.
- McDonald JH, Kreitman M (1991) Adaptive protein evolution at the *Adh* locus in *Drosophila*. *Nature* 351: 652–654.
- Mendel GJ (1866) Versuche über pflanzenhybriden. *Verhandlungen des naturforschenden Vereins Brünn* IV: 3–47.
- Miyata T, Yasunaga T (1980) Molecular evolution of mRNA: a method for estimating evolutionary rates of synonymous and amino acid substitutions from homologous nucleotide sequences and its application. *J. Mol. Evol.* 16: 23–36.
- Mohd-Sarip A, Venturini F, Chalkley GE, Verrijzer CP (2002) Pleiohomeotic can link Polycomb to DNA and mediate transcriptional repression. *Mol. Cell. Biol.* 22: 7473–7483.
- Monclús M (1976) Distribución y ecología de Drosophilidos en España. II. Especies de *Drosophila* de las Islas Canarias, con la descripción de una nueva especie. *Bol. Real Soc. Española Hist. Nat.* 74: 197–213.
- Monclús M (1984) Drosophilidae of Madeira, with the description of *Drosophila madeirensis* n. sp. *J. Zool. Syst. Evol. Res.* 22: 94–103.
- Müller J, Verrijzer P (2009) Biochemical mechanisms of gene regulation by Polycomb group protein complexes. *Curr. Opin. Genet. Dev.* 19: 150–158.
- Nekrasov M, Klymenko T, Fraterman S, Papp B, Oktaba K et al. (2007) Pcl-PRC2 is needed to generate high levels of H3-K27 trimethylation at Polycomb target genes. *EMBO J.* 26: 4078–4088.
- Nekrasov M, Wild B, Müller J (2005) Nucleosome binding and histone methyltransferase activity of *Drosophila* PRC2. *EMBO Rep.* 6: 348–353.
- Ng J, Li R, Morgan K, Simon J (1997) Evolutionary conservation and predicted structure of the *Drosophila* extra sex combs repressor protein. *Mol. Cell. Biol.* 17: 6663–6672.
- Nielsen R, Yang Z (1998) Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics* 148: 929–936.
- Obbard DJ, MacLennan J, Kim K-W, Rambaut A, O'Grady PM et al. (2012) Estimating divergence dates and substitution rates in the *Drosophila* phylogeny. *Mol. Biol. Evol.* 29: 3459–3473.

- Ohno K, McCabe D, Czermin B, Imhof A, Pirrotta V (2008) ESC, ESCL and their roles in Polycomb group mechanisms. *Mech. Dev.* 125: 527–541.
- Okulski H, Druck B, Bhalarao S, Ringrose L (2011) Quantitative analysis of Polycomb response elements (PREs) at identical genomic locations distinguishes contributions of PRE sequence and genomic environment. *Epigenetics Chromatin* 4: e4.
- Papacit M, San Antonio J, Prevosti A (1991) Genetic analysis of extra sex combs in the hybrids between *Drosophila subobscura* and *D. madeirensis*. *Genetica* 84: 107–114.
- Payer B, Lee JT (2008) X chromosome dosage compensation: how mammals keep the balance. *Annu. Rev. Genet.* 42: 733–772.
- Peel AD, Chipman AD, Akam M (2005) Arthropod segmentation: beyond the *Drosophila* paradigm. *Nat. Rev. Genet.* 6: 905–916.
- Peterson AJ, Kyba M, Bornemann D, Morgan K, Brock HW et al. (1997) A domain shared by the Polycomb group proteins Scm and ph mediates heterotypic and homotypic interactions. *Mol. Cell. Biol.* 17: 6683–6692.
- Pick L, Heffer A (2012) Hox gene evolution: multiple mechanisms contributing to evolutionary novelties. *Ann. N. Y. Acad. Sci.* 1256: 15–32.
- Pietersen AM, Van Lohuizen M (2008) Stem cell regulation by Polycomb repressors: postponing commitment. *Curr. Opin. Cell Biol.* 20: 201–207.
- Pratdesaba R, Segarra C, Aguadé M (2015) Inferring the demographic history of *Drosophila subobscura* from nucleotide variation at regions not affected by chromosomal inversions. *Mol. Ecol.* 24: 1729–1741.
- Pu L, Sung ZR (2015) PcG and trxG in plants – friends or foes. *Trends Genet.* 31: 252–262.
- Puerma E, Aguade M (2013) Polymorphism at genes involved in salt tolerance in *Arabidopsis thaliana* (Brassicaceae). *Am. J. Bot.* 100: 384–390.
- Quesada H, Ramos-Onsins SE, Aguadé M (2005) Birth-and-death evolution of the cecropin multigene family in *Drosophila*. *J. Mol. Evol.* 60: 1–11.
- Ramos-Onsins SE, Puerma E, Balaña-Alcaide D, Salguero D, Aguadé M (2008) Multilocus analysis of variation using a large empirical data set: phenylpropanoid pathway genes in *Arabidopsis thaliana*. *Mol. Ecol.* 17: 1211–1223.
- Randsholt NB, Santamaria P (2008) How *Drosophila* change their combs: the Hox gene *Sex combs reduced* and sex comb variation among *Sophophora* species. *Evol. Dev.* 10: 121–133.
- Remsen J, O'Grady P (2002) Phylogeny of Drosophilinae (Diptera: Drosophilidae), with comments on combined analysis and character support. *Mol. Phylogenet. Evol.* 24: 249–264.
- Richards S, Liu Y, Bettencourt BR, Hradecky P, Letovsky S et al. (2005) Comparative genome sequencing of *Drosophila pseudoobscura*: chromosomal, gene, and cis-element evolution. *Genome Res.* 15: 1–18.
- Robinson AK, Leal BZ, Chadwell LV, Wang R, Ilango U et al. (2012) The growth-suppressive function of the Polycomb group protein Polyhomeotic is mediated by polymerization of its sterile alpha motif (SAM) domain. *J. Biol. Chem.* 287: 8702–8713.

Bibliografía

- Roy S, Ernst J, Kharchenko PV, Kheradpour P, Negre N et al. (2010) Identification of functional elements and regulatory circuits by *Drosophila* modENCODE. *Science* 330: 1787–1797.
- Sánchez-Gracia A, Vieira FG, Rozas J (2009) Molecular evolution of the major chemosensory gene families in insects. *Heredity* 103: 208–216.
- Sanger F, Nicklen S, Coulson AR (1977) DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. U. S. A.* 74: 5463–5467.
- Sanger F, Tuppy H (1951) The amino-acid sequence in the phenylalanyl chain of insulin. I. The identification of lower peptides from partial hydrolysates. *Biochem. J.* 49: 463–481.
- Scheuermann JC, Gaytán De Ayala Alonso A, Oktaba K, Ly-Hartig N, McGinty RK et al. (2010) Histone H2A deubiquitinase activity of the Polycomb repressive complex PR-DUB. *Nature* 465: 243–247.
- Scheuermann JC, Gutiérrez L, Müller J (2012) Histone H2A monoubiquitination and Polycomb repression: the missing pieces of the puzzle. *Fly* 6: 162–168.
- Schuettengruber B, Ganapathi M, Leblanc B, Portoso M, Jaschek R et al. (2009) Functional anatomy of Polycomb and trithorax chromatin landscapes in *Drosophila* embryos. *PLoS Biol.* 7: e1000013.
- Schuettengruber B, Martinez AM, Iovino N, Cavalli G (2011) Trithorax group proteins: switching genes on and keeping them active. *Nat. Rev. Mol. Cell Biol.* 12: 799–814.
- Shao Z, Raible F, Mollaaghababa R, Guyon JR, Wu CT et al. (1999) Stabilization of chromatin structure by PRC1, a Polycomb complex. *Cell* 98: 37–46.
- Shaver S, Casas-Mollano JA, Cerny RL, Cerutti H (2010) Origin of the Polycomb repressive complex 2 and gene silencing by an *E(z)* homolog in the unicellular alga *Chlamydomonas*. *Epigenetics* 5: 301–312.
- Simon JA, Kingston RE (2009) Mechanisms of Polycomb gene silencing: knowns and unknowns. *Nat. Rev. Mol. Cell Biol.* 10: 697–708.
- Sinclair DA, Milne TA, Hodgson JW, Shellard J, Salinas CA et al. (1998) The Additional sex combs gene of *Drosophila* encodes a chromatin protein that binds to shared and unique Polycomb group sites on polytene chromosomes. *Development* 125: 1207–1216.
- Smith MD, Wertheim JO, Weaver S, Murrell B, Scheffler K et al. (2015) Less is more: an adaptive branch-site random effects model for efficient detection of episodic diversifying selection. *Mol. Biol. Evol.* 32: 1342–1353.
- Song JJ, Garlick JD, Kingston RE (2008) Structural basis of histone H4 recognition by p55. *Genes Dev.* 22: 1313–1318.
- Steffen PA, Ringrose L (2014) What are memories made of? How Polycomb and Trithorax proteins mediate epigenetic memory. *Nat. Rev. Mol. Cell Biol.* 15: 340–356.
- Stirnimann CU, Petsalaki E, Russell RB, Müller CW (2010) WD40 proteins propel cellular networks. *Trends Biochem. Sci.* 35: 565–574.
- Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 123: 585–595.

- Tanaka K, Barmina O, Sanders LE, Arbeitman MN, Kopp A (2011) Evolution of sex-specific traits through changes in HOX-dependent *doublesex* expression. *PLoS Biol.* 9: e1001131.
- Tie F, Stratton CA, Kurzhals RL, Harte PJ (2007) The N terminus of *Drosophila* ESC binds directly to histone H3 and is required for E(Z)-dependent trimethylation of H3 lysine 27. *Mol. Cell. Biol.* 27: 2014–2026.
- Van Bortle K, Ramos E, Takenaka N, Yang J, Wahi JE et al. (2012) *Drosophila* CTCF tandemly aligns with other insulator proteins at the borders of H3K27me3 domains. *Genome Res.* 22: 2176–2187.
- Wang L, Brown JL, Cao R, Zhang Y, Kassis JA et al. (2004) Hierarchical recruitment of Polycomb group silencing complexes. *Mol. Cell* 14: 637–646.
- Wang L, Jähren N, Miller EL, Ketel CS, Mallin DR et al. (2010) Comparative analysis of chromatin binding by Sex comb on midleg (SCM) and other Polycomb group repressors at a *Drosophila* Hox gene. *Mol. Cell. Biol.* 30: 2584–2593.
- Wang W, Thornton K, Berry A, Long M (2002) Nucleotide variation along the *Drosophila melanogaster* fourth chromosome. *Science* 295: 134–137.
- Wang Y, Hu X-J, Zou X-D, Wu X-H, Ye Z-Q et al. (2015) WDSPdb: a database for WD40-repeat proteins. *Nucleic Acids Res.* 43: D339–D344.
- Watson JD, Crick FHC (1953) Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid. *Nature* 171: 737–738.
- Wen P, Quan Z, Xi R (2012) The biological function of the WD40 repeat-containing protein *p55/Caf1* in *Drosophila*. *Dev. Dyn.* 241: 455–464.
- Wright S (1931) Evolution in mendelian populations. *Genetics* 16: 97–159.
- Wu H-A, Bernstein E (2008) Partners in imprinting: noncoding RNA and Polycomb group proteins. *Dev. Cell* 15: 637–638.
- Yang Z (2007) PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* 24: 1586–1591.
- Yang Z, Nielsen R (1998) Synonymous and nonsynonymous rate variation in nuclear genes of mammals. *J. Mol. Evol.* 46: 409–418.
- Yang Z, Swanson WJ (2002) Codon-substitution models to detect adaptive evolution that account for heterogeneous selective pressures among site classes. *Mol. Biol. Evol.* 19: 49–57.
- Yang Z, Wong WSW, Nielsen R (2005) Bayes empirical Bayes inference of amino acid sites under positive selection. *Mol. Biol. Evol.* 22: 1107–1118.
- Zhou Q, Bachtrog D (2012) Sex-specific adaptation drives early sex chromosome evolution in *Drosophila*. *Science* 337: 341–345.



GCTGTGTTTGCAGGGAAGTCCCTAGCTGCAAGGCAACGGGATACACAGCAGGCTTCTGTGCC
AAGGTGTTTCATCAAAAACAGAAACCCCAATAGGAGTAAACAGCAACAGGACCTTCTGTGCC
TAGGACCGGTCCGCTAAATGGGACATCTACGAGCAGAGCAACAAATCTGCTGTGCC
GCAAAAAGCAGAGCCCCCTGACTCCATTGTGTCGCCAAGGGCTTTTACCCTCTGCTGTGCC
ACATGCTGAGAAGAGACAAATGACTACTATAAGGAGAGTGTCTGACCTTCTGTGCC
ACGAACAGCACCCGGAGCCAGGCACATCTGGCGAGCCCGGACTTACAGGCTTGTGAAC
CCACAGGCCAATGCGGACTGTGGATCATCTCAGGCAAGATATCTGAGGCTTGTGAAC
GACGATGGTGAAGGGGGGGCGCGGGTTTCAGTGTGTGCGAGAGCTTGTGAAC
GGAGTTCTCGCACACCGCGGACTGGATCGCTTTGTGCGAGAGCTTGTGAAC
GAGAAGTTGCGCACCGCTCTATCTGACGGCTGGGAATACCTTGTGAAC
GATCGTCAATGTATCTGTACGATCTGCCCGCGCAGAAAGAGAAAGTGTGTGCC
AGTATCTGTCGCCCCAGAGCAAAATGCAATCTGGGCAAAATGCAAGTGTGTGCC
CCCAAGTGTGCGCCGAGAGCAAAAGCAACTGGGCCACCGTGGAGAGGTGTGCC
CGCACTGGGAAGCACGGGCATGAAGCGGCGACCATCAGCTTGTGAAGGTGTGCC
CATCTTTCAGGGTATCTGCGCGCAGGAGACGCTGGTCTACCTTGTGTGTGCC
CCGTCCGCGATATTCTGTGTCGCGGCCCAAGGACTCTGCGGCCCGG
ACACAGTCCGGTGTATACGATTTTGCAGAGCAAACTCCAAGCATGATGA