

# **Estimación de la huella de la selección natural y el efecto Hill-Robertson a lo largo del genoma de *Drosophila melanogaster***

*Estimació de la petjada de la selecció natural i l'efecte Hill-Robertson al genoma de Drosophila melanogaster*

*Estimating the fingerprint of natural selection and the Hill-Robertson effect along the genome of Drosophila melanogaster*

**TESIS DOCTORAL**

Doctorado en genética

**David Castellano Esteve**

Director: Antonio Barbadilla Prados



Universitat Autònoma de Barcelona  
Facultat de Biociències  
Departament de Genètica i de Microbiologia  
Bellaterra, abril 2016







*A Elia y Diana*



“el camino es siempre mejor que la posada”

Cervantes



## **AGRADECIMIENTOS**

---

Gracias a todos los que habéis compartido vuestro tiempo, conocimiento, amistad y amor durante esta etapa de mi vida.



# CONTENIDOS

---

<b><i>Summary</i></b> .....	1
<b>1 Introducción</b> .....	5
1.1 Genética de poblaciones: contextualización histórica .....	5
1.1.1 Los procesos evolutivos fundamentales y su modelización matemática.....	12
1.1.2 Teoría casi neutra de la evolución molecular .....	23
1.1.3 Efecto Hill-Robertson.....	34
1.2 De la teoría a la práctica: retos actuales de la genética de poblaciones .....	45
1.2.1 Demografía vs selección ligada.....	46
1.2.2 Paradoja de la variación genética.....	51
1.2.3 Tasa y tiempo de fijación de mutaciones adaptativas.....	59
1.3 Estimación de la huella de la selección a lo largo del genoma .....	65
1.3.1 Tipos de selección natural.....	69
1.3.2 Pruebas de la teoría neutralista ( <i>Neutrality-Tests</i> ) .....	75
1.3.3 Nuevas aproximaciones a las pruebas de neutralidad.....	81
1.4 <i>Drosophila melanogaster</i> .....	86
<b>1.5 Objetivos</b> .....	92
<b>2 Materiales y métodos</b> .....	95
2.1 Secuencias poblacionales genómicas .....	95
2.1.1 Secuencias proyecto DGRP .....	95
2.1.2 Líneas consanguíneas y variación natural.....	97
2.1.3 Cuellos de botella, estructura poblacional y variación natural .....	100
2.2 Tratamiento de los datos y cálculo de los estadísticos.....	101
2.2.1 Incorporación de las anotaciones génicas a los datos genómico poblacionales .....	101
2.2.2 Criterios aplicados para el filtrado de los datos .....	103
2.2.3 Estimación del espectro de frecuencias (SFS) y la divergencia .....	105
2.3 Simulaciones de los estimadores de la selección purificadora.....	108
2.3.1 Libre recombinación entre sitios y censo efectivo constante .....	108
2.3.2 Sitios ligados y censo efectivo variable .....	112
2.4 Estimación de la tasa de recombinación .....	116
2.5 Estimación de la densidad génica.....	118
2.6 Estimación de la tasa de mutación .....	119
2.7 Estimación de la <i>DFE</i> y la tasa de evolución adaptativa .....	121
2.8 Estimación del sesgo en el uso de codones .....	124
2.9 Genes del sistema inmune y expresión sesgada en machos.....	124
2.10 Análisis estadístico .....	125
2.10.1 Correlaciones, regresiones y ANCOVA.....	125
2.10.2 Estimación del efecto Hill-Robertson en el genoma.....	126

2.10.3 Prueba de permutación, <i>Bootstraps</i> , cálculo de intervalos de confianza y valores p .....	127
2.11 Software y scripts .....	130
<b>3 Resultados .....</b>	<b>131</b>
3.1 Definición de los nuevos estimadores de la selección negativa .....	131
3.1.1 Simulaciones con sitios que segregan libremente .....	134
3.1.2 Simulaciones con sitios ligados y demografía .....	138
3.1.3 Estimación del valor p de los estimadores $d_n$ y $b$ .....	142
3.2 Estimación de la huella de la selección natural a lo largo del genoma codificador y no-codificador de <i>Drosophila melanogaster</i> .....	144
3.2.1 Mutaciones deletéreas y efectivamente neutras .....	145
3.2.2 Mutaciones beneficiosas .....	151
3.2.3 Mutaciones efectivamente seleccionadas y recombinación .....	151
3.2.4 Cromosoma X vs autosomas .....	154
3.3 Cuantificación del efecto Hill-Robertson sobre las mutaciones beneficiosas de cambio de aminoácido en <i>Drosophila melanogaster</i> .....	159
3.3.1 Tasa de recombinación y adaptación .....	160
3.3.2 Densidad génica y adaptación .....	166
3.3.3 Tasa de mutación y adaptación .....	168
3.3.4 La proporción de mutaciones adaptativas no fijadas debido al efecto Hill-Robertson ....	172
<b>4 Discusión .....</b>	<b>177</b>
4.1 Estimadores $d_n$ y $b$ : nuevas pruebas estadísticas de la acción de la selección purificadora.....	178
4.2 DFE de las nuevas mutaciones deletéreas y tasa de evolución adaptativa en el genoma de <i>D. melanogaster</i> ..	183
4.2.1 Fracción funcional del genoma de <i>D. melanogaster</i> .....	184
4.2.2 Contribución de las mutaciones codificadoras y no-codificadoras a la variación en la eficacia biológica entre individuos de <i>D. melanogaster</i> .....	189
4.2.3 Contribución de las substituciones codificadoras y no-codificadoras a la evolución adaptativa en <i>Drosophila</i> .....	191
4.2.4 Evolución lenta del X: mutaciones deletéreas recessivas .....	197
4.2.5 Evolución rápida del X: mutaciones beneficiosas recessivas .....	202
4.2.6 Inversiones y composición génica diferencial entre brazos cromosómicos .....	206
4.3 Efecto Hill-Robertson sobre la tasa de evolución adaptativa en proteínas de <i>D. melanogaster</i> .....	210
4.3.1 ¿Es la tasa de adaptación proteica independiente de la tasa de mutación? .....	211
4.3.2 iHR y constancia de la DFE a lo largo del genoma .....	211
4.3.3 Escasa importancia de la conversión génica para la iHR .....	213
4.3.4 Posible impacto de la interferencia de Hill-Robertson en la adaptación humana .....	215
4.3.5 ¿Nuevo método para la inferencia de la DFE de nuevas mutaciones beneficiosas? .....	216
4.3.6 Selección sobre modificadores de la tasa de recombinación .....	216
<b>Conclusions .....</b>	<b>219</b>
<b>Bibliografía .....</b>	<b>221</b>
<b>ANEXO (acceso contenido online) .....</b>	<b>249</b>





## SUMMARY

---

The present thesis is a comprehensive population genomic study in the model species *Drosophila melanogaster* which combines bioinformatic and theoretical-statistical approaches. The purpose of this study is two-fold, we want to estimate the relative importance of different regimes of natural selection and the impact of the Hill-Robertson effect along the genome of *D. melanogaster*. We have divided the results and discussion sections in three distinct parts. In the first part we have designed two new point estimators of the action of purifying selection using nucleotide polymorphism data. In the second part we have applied our new estimators together with existing methods to estimate the distribution of fitness effects (*DFE*) of new deleterious mutations and the rate of adaptive evolution using genomic sequences of *D. melanogaster* coming from the *Drosophila* Genetic Reference Panel (DGRP) project. This work has started as part of a large international project which was published in 2012 (Mackay *et al.* 2012). In the third and last part we have studied the effect of the rate of recombination, the mutation rate and gene density upon the number of adaptive amino acid substitutions that have occurred since the split between *D. melanogaster* and *D. yakuba*. Moreover, we have estimated for the first time how much the Hill-Robertson interference impedes the rate of adaptive evolution in the coding genome of *D. melanogaster*. This work has been recently published in 2016 (Castellano *et al.* 2016).

Our new estimators (called  $d_n$  and  $b$ ) use the information contained in the site frequency spectrum of putatively selected and neutral sites to infer the action of recent negative selection. Although  $d_n$  and  $b$  are qualitative estimators of the underlying *DFE* of new deleterious mutations, our simulation study shows that they

both increase when the average strength of negative selection and/or the proportion of effectively deleterious mutations increases.  $d_n$  is robust to non-equilibrium conditions while  $b$  estimates are biased under recent demographic changes. Both estimators can be used with confidence if tens or hundreds of segregating sites are available.  $d_n$  and  $b$  estimates are highly correlated to two previous maximum likelihood estimates of the fraction of strongly and weakly selected new deleterious mutations, respectively.

We show that purifying selection is pervasive along the coding and non-coding genome of *D. melanogaster*. We estimate that ~ 60% of the *D. melanogaster* genome is under recent purifying selection (or functional) and the majority of slightly deleterious mutations occur on non-coding sequences. Despite of this, it is very likely that nonsynonymous mutations will explain most of the variance in fitness between individuals because the average strength of selection is higher for coding than non-coding mutations. Regarding adaptive evolution, the estimate of the rate of adaptive substitutions in non-coding sequences is on average two fold the estimate in coding regions (relative to the mutation rate). Interestingly, we do not find significant differences in the rate of adaptive evolution between non-coding class sites (UTR, introns or intergenic regions). Our analysis suggests that regulatory mutations might dominate the adaptive change since the split between *D. melanogaster* and *D. yakuba*.

We have also distinguished patterns of coding and non-coding evolution among chromosome arms. We find that the fraction of slightly deleterious nonsynonymous mutations is lower in the X chromosome than in the autosomes (after controlling for gene density and recombination rate differences). This seems to be due to the presence of (partially) recessive new deleterious mutations. We show that the rate of adaptive evolution is higher in the X chromosome than in the autosomes for both coding and non-coding mutations (after controlling for gene density and recombination rate differences). Again, this can be due to the presence of (partially)

recessive new beneficial mutations. These results support both the hypothesis that adaptation arises from new mutations and that of the faster X-effect. There are some evidences which indicate that the efficacy of selection (negative and positive) is lower in the chromosome arm 3R relative to the rest of chromosome arms. We discuss that this lower selection efficacy could be due to the presence of more polymorphic (and fixed) large inversions in this chromosome arm compared to the rest of (large) chromosome arms.

Our analysis show how the rate of recombination, the mutation rate and gene density affect the rate of adaptation within the *D. melanogaster* genome. We find that the rate of adaptive amino acid substitution is positively correlated to both recombination rate and an estimate of the mutation rate, while it is negatively correlated to gene density. We also find that this correlation is robust to controlling for each other, synonymous codon bias and gene functions related to immune response and testes. We find that the adaptation rate depends on the mutation rate, thus adaption seems to be mutation limited. We estimate that on average at least ~27% of all advantageous substitutions have been lost because of HRi and that this quantity depends on gene's mutation rate and the gene density where the gene is located: genes with low mutation rates embedded in gene poor regions lose ~17% of their adaptive substitutions while genes with high mutation rates embedded in gene rich regions lose ~60%. We show that recombination, mutation and gene density are important determinants of the rate of adaptive evolution within the *Drosophila* genome.



# INTRODUCCIÓN

---

## 1.1 GENÉTICA DE POBLACIONES: CONTEXTUALIZACIÓN HISTÓRICA

La genética de poblaciones es la ciencia encargada de describir e interpretar los cambios que ocurren en las frecuencias alélicas, o composición genética de las poblaciones naturales, generación tras generación (Dobzhansky 1937). A este tipo de cambios intra-poblacionales, o especie-específicos, también se les denomina cambios micro-evolutivos, en contra posición a los cambios macro-evolutivos que observamos entre especies. La distinción entre ambos tipos de cambios es una mera consecuencia de la escala temporal a la que observemos la evolución, pues esperamos que los cambios micro-evolutivos desemboquen a largo plazo en cambios macro-evolutivos. Tanto micro como macro-evolución se basan en los mismos procesos evolutivos fundamentales: mutación, selección natural, migración, estructura poblacional, deriva genética y recombinación. La genética de poblaciones proporciona en definitiva el marco teórico en el que fundamentar la evolución a nivel molecular.

G. H. Hardy y W. R. Weinberg desarrollaron de manera independiente en 1908 un principio que servía como modelo nulo para explicar el mantenimiento de las frecuencias alélicas y genotípicas de una generación a la siguiente en poblaciones panmícticas en ausencia de cualquier fuerza evolutiva. Aunque la sencillez matemática del principio de Hardy-Weinberg roza lo trivial, este fue muy importante. En la etapa pre-Mendeliana se creía que la variación genética se reducía a la mitad cada generación. Antes del descubrimiento, o redescubrimiento (por Vries, Correns y Tschermark de manera independiente en 1900), de la herencia de partículas discretas por Mendel (1866), denominadas más tarde genes, no había ningún mecanismo de segregación. Es más, se creía que la información genética de los progenitores estaba

mezclada en la sangre de estos (*blending inheritance*) y que la progenie mostraba siempre fenotipos intermedios a los de los progenitores. Esto suponía un fuerte contratiempo para la aceptación completa de la evolución por selección natural de Darwin, pues en ausencia de variación esta no podía operar. Esto llevó a Darwin a magnificar el papel de la mutación dirigida por el ambiente en posteriores ediciones de su obra sobre el origen de las especies con tal de mantener un flujo constante de variación y así la validez de su teoría. La aceptación por parte de la comunidad científica de la segregación mendeliana hizo que todos estos problemas se desvanecieran, pero ni Darwin ni Mendel vivieron este hecho.

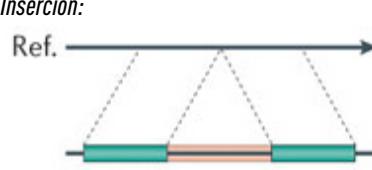
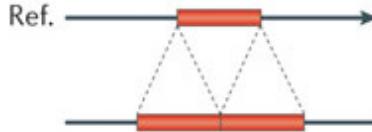
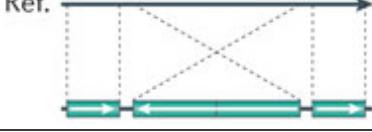
No obstante, los padres de esta ciencia tal y como la conocemos hoy día no fueron G. H. Hardy y W. R. Weinberg, sino R. A. Fisher, J. B. S. Haldane y S. Wright, que entre la segunda y tercera década del siglo XX calcularon las consecuencias del azar y la selección sobre la frecuencia de mutaciones independientes con herencia mendeliana, convirtiendo a la genética de poblaciones en el núcleo duro de la teoría de la evolución de Darwin. Al final de los años 30 y durante los años 40 se integró la genética de poblaciones teórica junto a la evolución experimental de poblaciones, la paleontología, sistemática, zoología y botánica en lo que se denominó la Síntesis Moderna de la biología evolutiva (Dobzhansky 1937; Mayr 1942; Simpson 1944; Stebbins 1950), denominada por algunos Neo-Darwinismo. Principalmente la Síntesis Moderna supuso la incorporación de la idea de gen, de las leyes de Mendel sobre la herencia, en la teoría original de la evolución por selección natural de Darwin.

A partir de entonces el debate se ha centrado en conocer la importancia relativa de cada proceso evolutivo (mutación, selección natural, migración, estructura poblacional, deriva genética y la más recientemente incorporada recombinación) en los niveles y patrones de variación genética que observamos en las poblaciones naturales. La segunda gran cuestión en evolución y en genética de poblaciones es identificar sobre qué caracteres o fenotipos actúa la selección natural y la relación

de estos con el genotipo. Resolver esta cuestión requiere conocer la relación entre el genotipo y el fenotipo, el mapa genotipo-fenotipo (Lewontin 1974), o la arquitectura génica de los caracteres, algo que para la mayoría de caracteres cuantitativos tenemos información muy limitada. La genética cuantitativa (Falconer 1960; Falconer y Mackay 1996) es la ciencia encargada de descifrar este mapa, sin embargo, la arquitectura génica de muchos caracteres cuantitativos, como la altura, es todavía un misterio aunque, aparentemente, estos caracteres sean altamente heredables (Zuk *et al.* 2012). Esta relación es más sencilla para los caracteres monogénicos o mendelianos. En la literatura encontramos múltiples ejemplos donde la acción de la selección natural se puede trazar de abajo a arriba, es decir, del genotipo (de los genes) al fenotipo, mediante una batería de test de selección, los cuales se pueden aplicar directamente sobre secuencias génicas o genómicas (véase sección 1.3).

No ha sido hasta fechas muy recientes que la genética de poblaciones no ha dispuesto de la cantidad masiva de genomas, tipos de mutaciones (tabla 1.1) y datos moleculares actuales (cuadro 1). El advenimiento de la electroforesis permitió estimar por primera vez los niveles de variación proteica que encontramos dentro de las poblaciones (Johnson *et al.* 1966; Lewontin y Hubby 1966, Harris 1966). A esta etapa se la denominó la ‘Era Alozímica’ (Lewontin 1974; 1992). Los geles electroforéticos de proteínas mostraron un tipo concreto de variación a nivel de ADN, las mutaciones puntuales no-sinónimas de cambio de aminoácido. La variación proveniente de mutaciones puntuales sinónimas y no-codificadoras, y el resto de mutaciones estructurales, permaneció oculta hasta finales de los 70 con la aparición de la secuenciación. [A excepción de grandes inversiones cromosómicas en *Drosophila*, que llevan siendo estudiadas desde principios del siglo pasado].

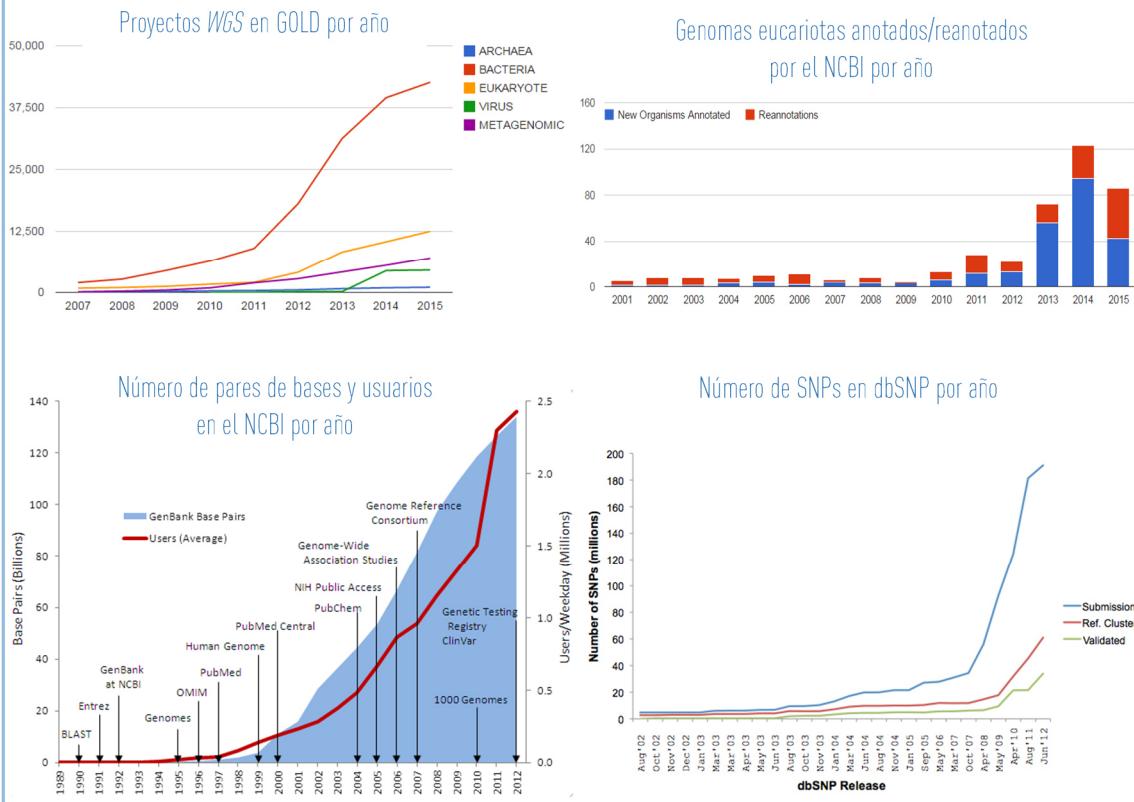
**TABLA 1.1 TIPOS DE MUTACIONES MÁS COMUNES EN EL ADN**

Tipo de variación	Descripción	
1. Mutaciones puntuales (SNV, <i>Single Nucleotide Variant</i> )	Substitución de bases que afecta a un solo nucleótido. Pueden ser transiciones o transversiones. Las mutaciones pueden ser codificadoras o no codificadoras.	ATGCAGTCGATCG <b>A</b> TGGCATGCATGC ATGCAGTCGATCG <b>C</b> TGGCATGCATGC
2. Insertiones y delecciones (INDEL)	Pares de bases extra que pueden añadirse (insertiones) o eliminarse (delecciones) del ADN.	<b>Delección:</b>  <b>Inserción:</b> 
3. Repeticiones en tandem de número variable (VNTR, <i>Variable number of tandem repeats</i> )	Un <i>locus</i> que contiene un número variable de ADN repetitivo en tandem que varía en longitud (2-8 nt para microsatélites, 7-100 nt para minisatélites) y es muy polimórfico.	
4. Variación en el número de copias (CNV, <i>Copy number variations</i> )	Una variante estructural genómica formada por copias de segmentos de $\geq 1$ kb (grandes duplicaciones).	
5. Duplicaciones Segmentales	Es un subtipo de CNV donde un par de fragmentos de $\geq 1$ kb tiene una identidad de secuencia $>90\%$ .	
6. Inversiones	Cambio en la orientación de un segmento de ADN.	
7. Translocaciones	Transferencia de un segmento de ADN a una región no homóloga del genoma. Normalmente es recíproca.	

[Tabla adaptada de Ràmia (2015)]

Pero previo a esto, o en paralelo, las enzimas de restricción fueron utilizadas también para estimar el polimorfismo en sus dianas de restricción a lo largo del genoma: el *Restriction Fragment Length Polymorphism* (o RFLP) fue la primera técnica de genotipación masiva debido a su bajo coste (Avise *et al.* 1983). Sin embargo, esta técnica estaba intrínsecamente limitada por el número de enzimas y dianas de restricción disponibles. El gran paso hacia adelante se produjo, sin duda, con la aparición de las tecnologías de secuenciación (Sanger y Coulson 1975, Maxam y Gilbert 1977), donde por fin toda la variación a nivel de ADN podía ser estimada directamente (Kreitman 1983). La automatización y parallelización del método de Sanger (Staden 1979; Smith *et al.* 1986; Mullis y Falloona 1987) ha permitido secuenciar muchos genomas, entre ellos el genoma humano (Lander *et al.* 2001; Venter *et al.* 2001). La secuenciación de Sanger reinó desde 1975 hasta 2005, con unas lecturas promedio de ~1000 bases, una tasa de error de 0,1 % y un coste aproximado de \$397 por megabase (coste mínimo en octubre de 2007, consultese evolución de costes en: <http://www.genome.gov/sequencingcosts/>) (Escalante *et al.* 2014). Actualmente el campo de la secuenciación se encuentra dominado por las tecnologías denominadas *next generation sequencing* (NGS). Centrándonos en la plataforma de NGS más común, vemos que proporciona unas lecturas promedio de 50-250 bases, una tasa de error de 0,26 % y un coste de \$0,05 por megabase (coste en abril de 2015) (Escalante *et al.* 2014).

## CUADRO 1: TSUNAMI DE GENOMAS Y VARIACIÓN GENÉTICA



En la última década hemos contemplado la generación y acumulación masiva de datos genómicos en multitud de repositorios públicos. El NCBI (*National Center for Biotechnology Information*), el cual forma parte de la biblioteca nacional de medicina de Estados Unidos, lidera el campo del almacenamiento de datos genómicos junto con el EMBL-EBI (*European Molecular Biology Laboratory – European Bioinformatics Institute*). El EMBL-EBI tiene una capacidad de almacenamiento de datos de unos 60 petabytes ( $1 \text{ PB} = 1 \times 10^{15}$  bytes), o el equivalente a 60 veces el genoma completo de todos los europeos. En el panel superior a la izquierda se muestra el número de WGS (*Whole Genome Sequence Projects*) depositados en el catálogo virtual de genomas GOLD (*Genomes OnLine Database*) por año y según el dominio de la vida al que pertenecen las especies. No obstante, esto corresponde sólo a genomas no anotados, en el panel superior derecho vemos como el número de especies con su genoma anotado y almacenado en el NCBI es significativamente menor. En el panel inferior izquierdo se muestra el número de pares de bases almacenados en el NCBI por año junto la aparición de los megaproyectos o consorcios internacionales y herramientas/plataformas bioinformáticas que han liderado dicha acumulación y análisis masiva de datos, respectivamente. Por último, en el panel inferior derecho muestro el número de SNPs (*Single Nucleotide Polymorphism*) almacenados en la base de datos pública dbSNP. La mayoría de estos SNPs ( $> 80\%$ ) provienen de la especie humana. [Imágenes extraídas de <https://gold.jgi.doe.gov/statistics> y <http://www.ncbi.nlm.nih.gov/>].

Podemos concluir que en la última década ha habido un abaratamiento espectacular de la secuenciación a costa de unas lecturas substancialmente más cortas y de ligera peor calidad. Esto ha supuesto que la cantidad de genomas completos y variantes genéticas detectadas hayan aumentado exponencialmente en la última década (cuadro 1). Antes de este tsunami de datos la genética de poblaciones era una ciencia principalmente teórica, el caldo de cultivo donde se propusieron distintas hipótesis y modelos evolutivos para explicar los niveles de variación genética que se esperaba encontrar en las poblaciones naturales. Estos niveles esperados de variación necesitaban ser contrastados y respaldados con datos reales con tal de conocer qué hipótesis podría empezar a denominarse teoría y qué hipótesis se verían relegadas a permanecer como elegantes modelos puramente teóricos.

En la sección 1.1.2 se explica por qué la teoría casi neutra de la evolución molecular de Ohta (Ohta 1972b; 1973; 1976; 1977; 1992) “ganó” la batalla y ha conseguido permear en todos los estudios de genética de poblaciones y evolución molecular más actuales, y por qué otras hipótesis han perdido protagonismo frente a ella. En la siguiente sección se explicará qué modelos y fenómenos genético-poblacionales son más importantes para enmarcar conceptualmente esta tesis. Finalmente, el efecto Hill-Robertson se describe en la sección 1.1.3.

### 1.1.1 LOS PROCESOS EVOLUTIVOS FUNDAMENTALES Y SU MODELIZACIÓN MATEMÁTICA

La genética de poblaciones es el campo de la biología con más teoría matemática. La razón principal por la cual la teoría se aplica más fácilmente a la genética de poblaciones es que hay un marco preciso - la segregación mendeliana - que la sustenta. La segregación mendeliana es altamente regular y de ella se derivan propiedades geométricas y algebraicas. Otro importante supuesto para la genética de poblaciones es el apareamiento al azar (*random mating*) o el supuesto de panmixia. Aunque no es un supuesto 100% cierto en la mayoría de casos, es un buen supuesto, pues muchos aspectos de la evolución no dependen de quién se aparee con quién y simplifica mucho el tratamiento matemático.

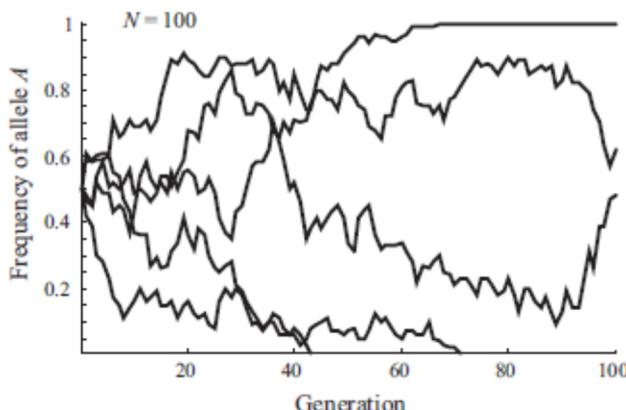
Muchos artículos en genética de poblaciones y evolución contienen modelos que raras veces se describen en detalle, además tienen descripciones incompletas de los supuestos y métodos utilizados. Como biólogo evolutivo es por lo tanto crucial saber “leer” las ecuaciones con tal de descifrar los supuestos que están implícitos en ellas. Esto no es sólo importante para entender mejor el artículo en concreto, es una habilidad indispensable para asegurarse que el modelo realmente describe el proceso para el cual ha sido diseñado. Además, los modelos suelen poder aplicarse a un conjunto de cuestiones mayor que el utilizado en un trabajo en concreto. En definitiva, es importante entender el significado de las ecuaciones para saber cuándo es seguro reutilizarlas en otras cuestiones enteramente distintas a las originales y saber si cumplen los requisitos necesarios para aplicarlas al problema original. En esta sección se describe el modelo de Wright-Fisher, la deriva genética y el censo efectivo ( $N_e$ ). Finalmente, se muestra el porqué de la importancia que ha tenido y tiene la teoría de la coalescencia en los estudios de evolución molecular.

## MODELO DE WRIGHT-FISHER: DEFINICIÓN DE LA DERIVA GENÉTICA

Consideremos una población (haploide o diploide hermafrodita) de tamaño constante,  $N$ , y sólo dos tipos de individuos, o alelos, ( $A$  y  $a$ ) con generaciones discretas (no solapantes); todos los individuos nacen, se reproducen y mueren a la vez. En este modelo los nuevos individuos son creados cada generación muestreando al azar con reemplazamiento los gametos producidos por la generación parental, la cual muere inmediatamente después de la reproducción. Cada progenitor tiene la misma probabilidad de contribuir a la siguiente generación. Como veremos este tipo de muestreo causa fluctuaciones al azar en las frecuencias alélicas respecto a la frecuencia esperada bajo el modelo determinístico. A esta variación en la frecuencia observada respecto a la esperada por el modelo determinístico es lo que denominamos formalmente como **deriva genética** (figura 1.1). El modelo determinístico de este proceso predice que la frecuencia del individuo (o alelo)  $A$  a tiempo  $t+1$  será exactamente

$$p(t+1) = W_A p(t) / (W_A p(t) + W_a (1 - p(t))) \quad (1.1)$$

En este modelo el número de individuos supervivientes por progenitor son  $W_A$  y  $W_a$ . Cuando  $W_A = W_a = 1$ , el denominado supuesto de neutralidad, la frecuencia de  $A$  y  $a$  debe permanecer constante ( $p(t+1) = p(t)$ ). No obstante, si para construir la población actual muestreamos al azar con reemplazamiento la población parental esperamos que por azar individuos de tipo  $A$  (o  $a$ ) puedan dejar más o menos descendencia en cualquier generación. Esto implica que el número de copias de  $A$  en la descendencia estará distribuido binomialmente con una media  $N p(0)$  y varianza  $p(1-p)/N$ , donde  $p(0)$  es la frecuencia inicial de  $A$ . Por lo tanto, para  $N$ s grandes la varianza en la frecuencia del alelo  $A$  será menor que bajo  $N$ s menores. Si el resultado  $j$  es el número de copias de  $A$  en la siguiente generación,  $p(1) = j/N$  y la frecuencia inicial  $p(0) = 1/2$  en una población de tamaño  $N = 100$ , en la siguiente generación obtenemos (por azar) 42 copias de  $A$ , y por lo tanto,  $p(1) = 0,42$ .



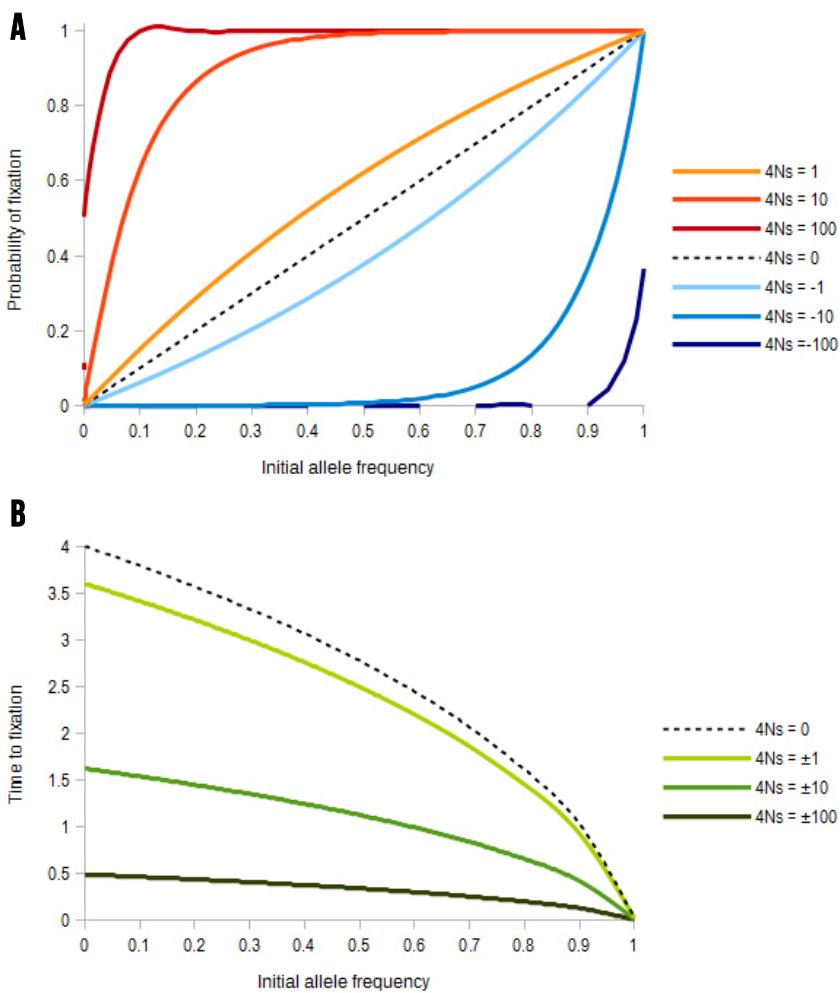
**FIGURA 1.1** El modelo de Wright-Fisher sin selección. Cada generación la descendencia es elegida al azar con reemplazamiento (esto es equivalente a un muestreo binomial). Asumimos que la población es de tamaño constante con  $N = 100$ . La frecuencia del alelo o individuos tipo  $A$  se representa a lo largo del tiempo, frecuencia inicial  $p(0) = 0.5$ . La **deriva genética** es dicha fluctuación estocástica observada respecto al modelo determinístico, donde la frecuencia se espera permanezca inalterada al 50%. [Figura tomada de Otto y Day (2007)].

Si iteramos este proceso 100 veces (o generaciones) y lo repetimos 5 veces de manera independiente obtenemos los resultados representados en la figura 1.1. Podemos usar simulaciones (como las realizadas en la figura 1.1), o matrices de transición, para determinar la probabilidad de que en una generación futura  $t$  la población esté compuesta por una proporción  $p(t)$  de individuos tipo  $A$ . Además, podemos calcular la probabilidad última de fijación de un alelo neutro dependiendo de su frecuencia inicial,  $i/N$ . Al hacer esto nos toparemos con un resultado clásico en genética de poblaciones: la probabilidad de fijación de un alelo neutro depende de su frecuencia inicial,  $i/N$ , y su probabilidad de extinción de  $1 - i/N$  (Kimura 1983).

La aproximación de difusión para la probabilidad de fijación de un alelo selectivo  $u(N, s)$  es otro de los grandes resultados en biología evolutiva y viene definido por esta expresión (Kimura 1957; 1983)

$$u(N, s) \sim \frac{1-e^{-2PNs}p_0}{1-e^{-2PNs}} \quad (1.2)$$

donde  $P$  es la ploidía ( $P = 1$  para haploides y  $P = 2$  para diploides),  $p_0$  es la frecuencia inicial,  $N$  el tamaño de población y  $s$  el coeficiente de selección. En la figura 1.2A se muestra la probabilidad de fijación dependiendo del coeficiente de selección, el tamaño poblacional y la frecuencia inicial. La figura 1.2B muestra el número de generaciones promedio que son necesarias para fijar mutaciones neutras en una población diploide (línea punteada) (Kimura y Ohta 1969). Según muestra la figura 1.2 cuando  $-1 < 4Ns < 1$  se espera que las mutaciones se comporten prácticamente como si fuesen neutras en lo que respecta a su dinámica poblacional: probabilidad de fijación y tiempo promedio para fijarse. Este tipo de mutaciones casi neutras o efectivamente neutras denominadas así por primera vez por Kimura (1968) son muy importantes para explicar los patrones de polimorfismo y divergencia que observamos en las poblaciones naturales (véase sección 1.1.2). A partir de  $|4Ns| > 1$  la selección empieza a tener un impacto significativo sobre la probabilidad y el tiempo de fijación, a estas mutaciones se las denomina comúnmente como ligeramente o débilmente seleccionadas (*slightly o weakly selected mutations*). A partir de  $|4Ns| > 10$  la selección ejerce un poder dramático sobre la dinámica poblacional de las mutaciones, a estas mutaciones se las denomina comúnmente como fuertemente seleccionadas (*strongly selected*). Son por tanto los coeficientes de selección escalados al tamaño poblacional los que tienen relevancia evolutiva, pues el mismo coeficiente de selección puede comportarse como deletéreo bajo un tamaño poblacional grande o como efectivamente neutro en un tamaño poblacional menor (véase sección 1.1.2).



**FIGURA 1.2** Probabilidad de fijación (A) y tiempo promedio para fijarse (B) según el modelo de Wright-Fisher con selección utilizando la aproximación de ecuaciones de difusión. Los valores de  $4Ns$  se muestran en la leyenda. (A) La probabilidad de fijación de un alelo neutro es igual a su frecuencia inicial tal y como mostramos anteriormente. (B) Sorprendentemente, el tiempo necesario para que un alelo deletéreo se fije es el mismo que el tiempo necesario para que un alelo beneficioso se fije (asumiendo la misma magnitud de la selección) (Ewens 1979). A primera vista este resultado parece anti-intuitivo, sin embargo, la razón es porque debe haber un cambio grande y rápido en la frecuencia alélica del alelo desfavorable para que este pueda fijarse. Estos cambios drásticos en la frecuencia son muy poco probables por eso la probabilidad de fijación es tan baja.

Una última derivación del modelo de Wright-Fisher es la distribución estacionaria de frecuencias. Wright derivó dos ecuaciones para describir la distribución de frecuencias alélicas: la reversible y la irreversible. La ecuación irreversible describe el tiempo que una mutación  $i$  pasa a frecuencia  $x$  (Wright 1938a). Esta ecuación ha sido tremadamente útil en este trabajo de tesis doctoral pues ha permitido calcular el espectro de frecuencias ante distintas distribuciones de coeficientes de selección (véase sección 2.3.1). La ecuación irreversible (Wright 1938)

$$\varphi(x) = \frac{(1-e^{-4Ne^s(1-x)})}{(1-e^{-4Ne^s})x(1-x)} \quad (1.3)$$

donde  $s$  es la fuerza de la selección actuando sobre la mutación.  $\varphi(x)$  es el tiempo que una nueva mutación pasa entre la frecuencia  $x$  y la frecuencia  $x + dx$ . La ecuación reversible en cambio describe la distribución de frecuencias alélicas de un alelo con mutación recurrente y selección (Wright 1929; 1931; 1937).

### CENSO EFECTIVO ( $N_e$ )

Hasta ahora se ha asumido que el tamaño de población  $N$  corresponde al número de individuos que contamos en una población. Esta  $N$  es de algún modo una medida de la intensidad de la deriva genética. En las poblaciones naturales sabemos que puede haber diferente número de machos y de hembras, las generaciones pueden solaparse, algunas especies (de plantas sobre todo) se autofecundan, hay casos de monogamia estricta, de variación en el tamaño de la población de forma estacional o por catástrofes naturales, hay también variación en el número de descendientes por individuo (por causa genética o ambiental), etc. La validez del modelo de Wright-Fisher, y sus predicciones, deben cuestionarse cuando sus supuestos se incumplen en las poblaciones naturales, de hecho, lo más normal es que se incumplan. ¿Podemos corregir todo esto de algún modo? En el modelo de Wright-Fisher el número de copias de  $A$  en la siguiente generación está distribuido binomialmente con una varianza  $p(1-p)/N$  en una población haploide (y  $p(1-p)/2N$  en una población

diploide). Hay que entender  $N_e$  como el número de individuos que satisface la varianza esperada en las frecuencias alélicas bajo la distribución binomial, y no como el número de individuos que contamos en una población, el denominado censo de población ( $N_c$ ). El concepto de  $N_e$  lo introdujo Wright por primera vez (1931, 1938b) como un intento de parametrizar la dinámica de un proceso poblacional complicado bajo un modelo evolutivo idealizado (este es, el modelo de Wright–Fisher). Hoy en día es un concepto omnipresente en cualquier trabajo de genética de poblaciones.

Todo proceso (genético o ambiental) que aumente la varianza en el número de descendientes por progenitor respecto a la varianza esperada por azar (esto es bajo la distribución binomial) acabará reduciendo  $N_e$  respecto  $N_c$ . A este efecto se le puede encontrar titulado en los libros de texto como el efecto del tamaño familiar, del número de gametos o de la variación en la *fitness*. La siguiente ecuación encapsula este efecto para una población de tamaño constante (Crow 1954)

$$N_e = \frac{4N_c - 2}{2 + V_n} \quad (1.4)$$

donde  $V_n$  es la varianza en el número de descendientes por progenitor. A medida que  $V_n$  aumenta  $N_e$  disminuye. En un caso extremo donde todos los individuos contribuyen por igual a la descendencia  $V_n = 0$ , por lo tanto,  $N_e = 2N_c - 1$ . En este caso  $N_e$  será prácticamente el doble que  $N_c$  (por eso en los programas de conservación de especies se intenta que los pocos individuos restantes se reproduzcan todos por igual). En el modelo de Wright–Fisher la población parental es muestreada al azar  $2N_c$  veces, en cada muestreo cada progenitor tiene  $1/N_c$  probabilidades de ser elegido para contribuir a la siguiente generación y  $1-1/N_c$  de no serlo. La varianza esperada en el número de descendientes de cada individuo viene descrita por la distribución binomial y es

$$V_n = 2N_c \left( \frac{1}{N_c} \right) \left( 1 - \frac{1}{N_c} \right) = 2 - \frac{2}{N_c}. \quad (1.5)$$

Substituyendo (1.5) en (1.4) obtenemos, obviamente, que  $N_e \sim N_c$ . Es importante resaltar que la selección natural, la cual por definición aumenta la varianza en el número de descendientes por progenitor, actúa también disminuyendo  $N_e$  respecto  $N_c$ . Este es sólo un ejemplo ilustrativo de la complejidad de la relación entre  $N_e$  y  $N_c$ , para una revisión extensa de la relación entre  $N_e$  y  $N_c$  consultese Caballero (1994) y Wang y Caballero (1999). En conclusión, en genética de poblaciones, la varianza de las frecuencias alélicas entre generaciones depende del número efectivo de individuos (estos son los individuos que acaban contribuyendo a la descendencia o el  $N_e$ ) y no del número total de individuos que contamos en una población o  $N_c$  (Charlesworth 2009). No obstante, el  $N_c$  actual (por si solo) puede tener un importante papel en la adaptación como se muestra en cuadro 4 (sección 1.3.1).

## IMPORTANCIA PRÁCTICA DE LA TEORÍA DE LA COALESCENCIA

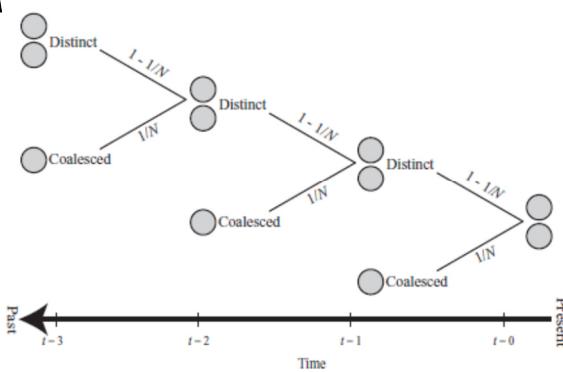
La genética de poblaciones clásica desarrollada como se ha comentado anteriormente en los años 30 y 40 del siglo pasado por R. A. Fisher, J. B. S. Haldane, S. Wright y ampliada más tarde en los 50 por M. Kimura y a partir de los 60 por T. Ohta (véase siguiente sección) era prospectiva, es decir, mostraba cómo la evolución operaba, o podía operar, a nivel de población partiendo de un punto del presente hacia el futuro, o de un punto del pasado hasta el presente. Es decir, siempre contemplaron, o mejor dicho modelaron, la evolución hacia adelante (*forward in time*). Todo esto cambió cuando los datos moleculares llegaron (veáse cuadro 1) y se empezó a disponer de *muestras*; alineamientos de ADN. Llegados a este punto las preguntas fueron otras. Dado un alineamiento de ADN ¿qué procesos evolutivos han ocurrido hasta llegar a él? ¿Ha jugado la selección natural algún papel? A principios de los años 80 J. Kingman (1982a; 1982b) desarrolló la teoría de la coalescencia; una teoría estocástica retrospectiva que operaba a nivel de muestras, no de poblaciones, y del presente hacia el pasado (*backward in time*) (Ewens 1979) (cuadro 2). Dado un modelo evolutivo estocástico (sin selección) la teoría de la coalescencia permite estimar los

niveles y patrones esperados de variación genética en una muestra de tamaño  $n$ . Proporciona una hipótesis nula con la que contrastar nuestros datos. No obstante, es difícil extender la teoría de la coalescencia a escenarios con múltiples *loci* bajo selección (y recombinación) (véase Krone y Neuhauser 1997; Neuhauser y Krone 1997). A pesar de esta limitación, *software* tipo LAMARC basado en la teoría de la coalescencia (Kuhner 2006), puede recibir cantidades masivas de información proveniente de las secuencias de ADN y generar estimas de parámetros importantes como la tasa de mutación, recombinación y de migración (pero no estima la selección). La estimación integradora de todos estos parámetros junto con la selección bajo el marco teórico de la coalescencia puede estar más cerca gracias a los avances en la inferencia de parámetros mediante computación bayesiana aproximada (Beaumont *et al.* 2002) y/o algoritmos de aprendizaje automático (Jones 2014) (véase sección 1.3.3). No obstante, todavía desconocemos las ventajas asociadas a la construcción de modelos cada vez más complejos, pues estos conllevan un gran coste computacional cuando tratamos de inferir los parámetros mediante máxima verosimilitud. La pregunta clave es ¿podemos obtener respuestas igual de válidas mediante modelos simplificados de coalescencia e inferencia aproximada de parámetros que bajo modelos extremadamente realistas e inferencia por máxima verosimilitud (Wakeley 2009)?

En la actualidad no tenemos tan solo secuencias fragmentarias del genoma, sino genomas completos. Esto ha supuesto, creo, no una nueva reformulación de nuestras preguntas, seguimos queriendo saber qué procesos evolutivos han ocurrido hasta dar lugar a nuestros genes y genomas. Pero sí ha supuesto la búsqueda y el diseño de nuevas herramientas con tal de poder responderlas. En definitiva, aunque computacionalmente las simulaciones mediante coalescencia son más eficaces que las simulaciones *forward in time*, estas últimas tienen la ventaja de ser más flexibles pues permiten simular los efectos de la selección en presencia del resto de fuerzas evolutivas (véase sección 1.3.3).

## CUADRO 2: VIAJANDO AL PASADO: TEORÍA DE LA COALESCENCIA

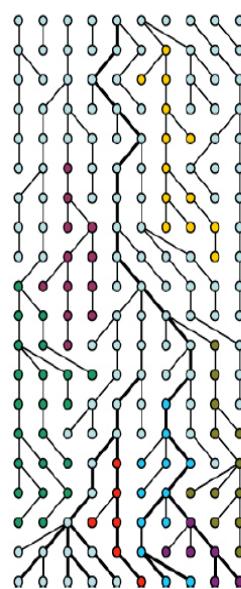
A



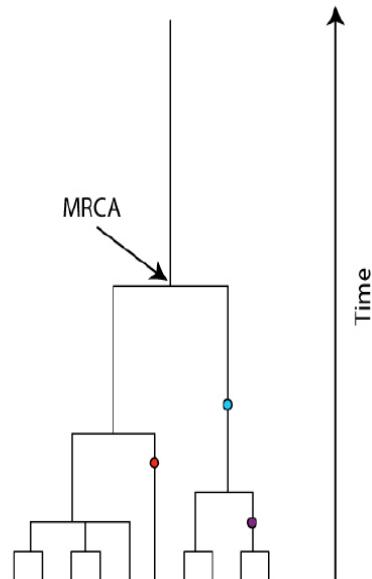
Dentro de los modelos dinámicos siempre imaginamos que el tiempo ocurre hacia adelante (*forward in time*). La teoría de la coalescencia (Kingman 1982a; 1982b) describe la probabilidad de que los alelos presentes en una muestra actual desciendan de un mismo alelo ancestral  $t$  generaciones en el pasado y puede representarse como un árbol de decisión (A). No deja de ser un modelo de tipo estocástico, pues hace los mismos supuestos que el modelo de Wright-Fisher, pero hacia atrás en el tiempo (*backward in time*).

El término coalescencia se refiere al proceso en el que mirando hacia atrás en el tiempo las genealogías de los alelos se unen en un ancestro común. Es decir, mediante este modelo se intenta relacionar todos los alelos de un gen compartidos por todos los individuos de una población a una única copia ancestral conocida como el ancestro común más cercano (o *most recent common ancestor, MRCA*). Una de las razones por las que el razonamiento de la coalescencia es tan poderoso es que el coste computacional 'hacia atrás' en el tiempo es mucho menor que si se analiza desde una perspectiva 'hacia delante' en el tiempo, en el que muchos alelos simulados son un cálculo desperdiciado ya que no dejan descendientes en el presente.

B



C



## CUADRO 2: CONTINUACIÓN

En el caso de (B), 164 de los 200 alelos simulados no aportarían información. Mientras que yendo hacia atrás en el tiempo sólo tendríamos que simular 36 alelos. Si imaginamos una muestra de cientos de alelos, el seguimiento de la mayoría de los linajes supone un coste innecesario.

Las relaciones de herencia entre alelos se representan típicamente mediante un árbol de coalecencia, o "genealogía génica". muy similar a un árbol filogenético (C). Comprender las propiedades estadísticas de estas genealogías bajo diferentes supuestos constituye el núcleo de la teoría de la coalecencia. Siguiendo en (C) podemos ver que yendo de la generación 0 (presente) a la generación anterior (1) hay cuatro sucesos de coalecencia, es decir, que 8 alelos actuales "coalecen" de dos en dos en un cuatro únicos alelos ancestral. A medida que vamos hacia atrás en el tiempo el número de alelos ancestrales se mantiene constante o disminuye y cada reducción en el número de alelos ancestrales es un suceso de coalecencia. Debido a que la mutación puede ocurrir durante este proceso, los alelos observados en el presente no serán idénticos en secuencia a los alelos originales, pero serán descendientes de un alelo ancestral común único. Bajo el modelo de coalecencia, el tiempo normalmente se mide en unidades de  $2N_e$  generaciones.

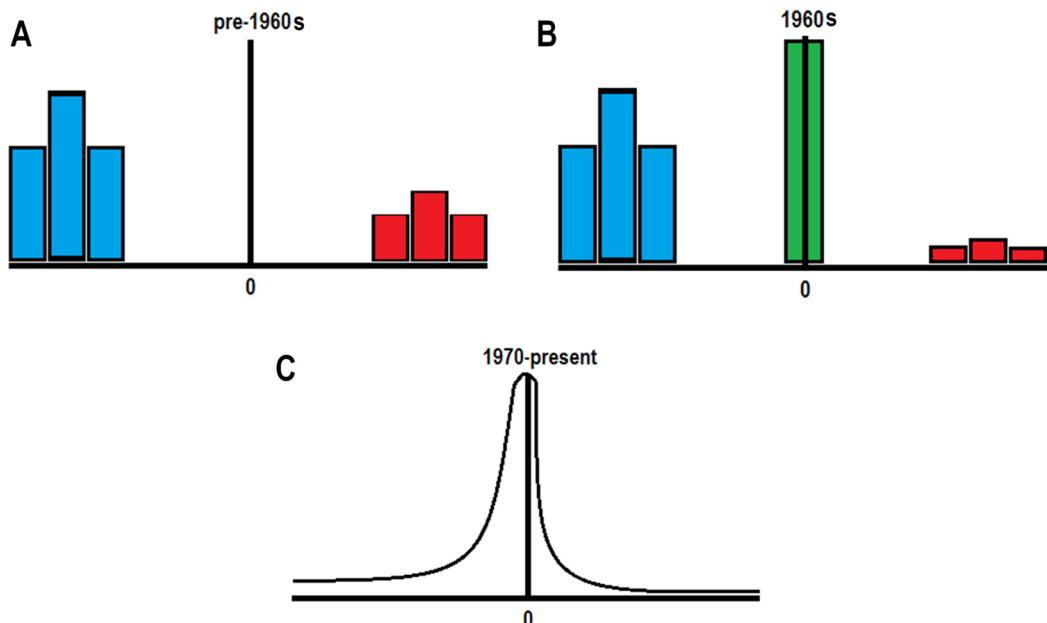
## 1.1.2 TEORÍA CASI NEUTRA DE LA EVOLUCIÓN MOLECULAR

Definir la teoría casi neutra de la evolución molecular requiere conocer la historia de otro concepto fundamental en genética de poblaciones y evolución: La distribución de los coeficientes de selección de las nuevas mutaciones (*Distribution of Fitness Effects, DFE*) (Eyre-Walker y Keightley 2007; Keightley y Eyre-Walker 2010).

El efecto que tiene una nueva mutación sobre la eficacia biológica del individuo se puede clasificar a grandes rasgos en tres categorías: deletérea, neutra o beneficiosa. No obstante, en realidad es más probable que las mutaciones sigan algún tipo de distribución continua, desde las mutaciones letales, fuertemente deletéreas, ligeramente deletéreas, pasando por las neutras, ligeramente beneficiosas hasta las fuertemente beneficiosas. Esta distribución de efectos sobre la eficacia biológica se le conoce formalmente como la *DFE*. La naturaleza de la *DFE* es un problema elemental en biología evolutiva, porque reside en el corazón de muchas otras cuestiones importantes. Se ha estado debatiendo durante 30 años en el debate neutralista-selecciónista (Gillespie 1991; Kimura 1983), pero también en muchos otros problemas en genética, como el mantenimiento de la variación genética y fenotípica cuantitativa (Maynard Smith y Haigh 1974; Charlesworth *et al.* 1993; 1995), la evolución del sexo y la recombinación (Peck *et al.* 1997), las consecuencias que acarrea un tamaño poblacional pequeño (Schultz y Lynch 1997), el reloj molecular (Ohta 1977; Kimura 1979; Ohta 1992) y la tasa de degradación genómica debido al trinquete de Muller (*Muller's ratchet*) (Loewe 2006).

La figura 1.3 muestra la evolución histórica de nuestra concepción de la *DFE* y con ella la teoría imperante en evolución molecular. ¿Qué ha ocurrido desde el pano-selecciónismo (figura 1.3A) de antes de los 60 del siglo pasado, donde la *DFE* era bimodal, hasta la *DFE* (de forma desconocida y que asumimos continua) (figura 1.3C) de hoy en día? A parte de indudables avances en la aplicación de herramientas matemáticas (ecuaciones de difusión, teoría de la coalescencia, etc.) y estadístico-

computacionales (inferencia de parámetros mediante máxima verosimilitud, computación bayesiana aproximada, desarrollo de simuladores *forward in time*), principalmente lo que ha ocurrido ha sido la llegada de nuevos datos que (bien interpretados) refutaban predicciones de teorías pretéritas.



**FIGURA 1.3** Evolución histórica de la DFE. En el eje de la x encontramos los coeficientes de selección. (A) Teoría pan-seleccionista de la evolución molecular. Todas las diferencias dentro de especies son o bien adaptativas (selección equilibradora) (esto es el *balance model* defendido por Dobzhansky 1970 y Ford 1971) o el resultado de un equilibrio entre la mutación y la selección negativa (esto es el *classical model* defendido por Muller y Kaplan 1966). (B) Teoría neutra de la evolución molecular (Kimura 1969a). La mayoría de diferencias en el ADN entre especies y dentro de especies son neutras, algunas diferencias entre especies son adaptativas (fijadas mediante selección positiva). (C) Teoría casi neutra de la evolución molecular (Ohta y Kimura 1971). La mayoría de diferencias en el ADN entre especies y dentro de especies son neutras, casi neutras y/o ligeramente deletéreas y algunas diferencias entre especies son adaptativas (fijadas mediante selección positiva). La teoría casi neutra debe verse como una versión más flexible de la teoría neutra.

Por ejemplo, en la década de 1960 el trabajo de Zuckerkandl y Pauling (1965) mostró que las hemoglobinas de mamíferos evolucionaban a una tasa constante de  $1,4 \times 10^{-7}$  substituciones aminoacídicas por año. Esto equivalía a la existencia de algo así

como un reloj molecular proteico (Zuckerkandl y Pauling 1965) que podía ser utilizado para calibrar las distancias evolutivas entre especies. Kimura (1968) fue el primero en percibir la importancia evolutiva de este hecho, utilizando la tasa de substitución estimada por Zuckerkandl y Pauling (1965) estimó que en el genoma de mamíferos “un nucleótido ha sido substituido en la población aproximadamente cada dos años”. Kimura cogió esta cifra y la utilizó para calcular el número de muertes por causa genética de acuerdo al trabajo de Haldane (1957); el valor que obtuvo era demasiado alto como para ser posible bajo las teorías pan-selecciónistas. A parte de datos de variación proteica entre especies, en esa misma década empezaron a llegar las primeras estimaciones de los niveles de variación proteica dentro de especies gracias a los geles de electroforesis (Johnson *et al.* 1966; Lewontin y Hubby 1966; Hubby y Lewontin 1966, Harris 1966). Los niveles de variación eran demasiado elevados para el modelo clásico de Muller y Kaplan (1966) la cual predecía que la mayoría de *loci* debían ser homocigotos. No obstante, en el modelo equilibrado (Dobzhansky 1970; Ford 1971) donde la selección equilibradora es ubicua resistía bien este nuevo dato pues precisamente se esperaba encontrar niveles elevados de variación genética para explicar la alta velocidad de adaptación observada en especies de *Drosophila*.

Las evidencias basadas en el coste o lastre genético de las mutaciones (*substitutional* o *genetic load*), fueron las que llevaron a Kimura a proponer que la mayoría de las mutaciones no-sinónimas de cambio de aminoácido debían ser neutras (Kimura 1969a) (figura 1.3B) (aunque años más tarde el propio Kimura reconoció que el argumento del lastre genético no era decisivo). En paralelo y de manera independiente King y Jukes (1969) publicaron otro artículo donde proponían prácticamente lo mismo pero con una mayor base bioquímica (pues King era bioquímico y Jukes un genético de poblaciones).

La teoría neutra incorporaba ideas interesantes:

- Las mutaciones deletéreas son rápidamente eliminadas de la población, y las mutaciones adaptativas son fijadas rápidamente. Por lo tanto, toda la variación dentro de especies debe ser selectivamente neutra (esto es que el alelo derivado tiene la misma eficacia biológica que el alelo ancestral o *wild-type*).
- El polimorfismo es una fase transitoria de la evolución molecular entre la extinción y la fijación, no se encuentra equilibrado por la selección natural.
- Los niveles de polimorfismo neutro ( $\Theta$ ) son el producto de la tasa de mutación neutra y el tamaño efectivo,  $N_e$ . Poblaciones grandes tendrán más polimorfismo que poblaciones pequeñas (no obstante, consultese la paradoja de los niveles de variación genética en la sección 1.2.2 [Lewontin 1974]).
- Las mutaciones neutras se fijan a una tasa constante ( $K$ ) que equivale al producto de la tasa de mutación por generación ( $\mu$ ) y la proporción de nuevas mutaciones neutras ( $f$ ),  $K = f \mu$ .

En la década de 1970 más información sobre la tasa de substituciones aminoacídicas en distintas proteínas se fue acumulando. Sin embargo, cada proteína tenía su tasa de substitución específica, desde las más rápidas, fibronopéptido a  $9 \times 10^{-9}$  substituciones por sitio y por año, a las más lentas como la histona IV a  $10 \times 10^{-11}$  por sitio y por año (Dickerson, 1971). Es decir, no había un único reloj molecular, sino múltiples, tantos como proteínas (consultese revisión de Kumar 2005). Kimura y Ohta (1974) interpretaron esta variación entre relojes moleculares del siguiente modo: las proteínas menos importantes, o las partes menos importantes de una proteína para su función, evolucionan más rápidamente, mientras que otras partes o proteínas más importantes están constreñidas selectivamente (o *selectively constrained*), estas no pueden cambiar pues reducen drásticamente la *fitness*.

La teoría neutralista fue ganando peso poco a poco por una sencilla razón – servía de modelo nulo con el que contrastar el modelo alternativo donde la selección natural sí tenía cabida. A igualdad de evidencias las hipótesis no-adaptacionistas son en general más parsimoniosas (Lynch 2007). La teoría neutralista original tenía dos talones de Aquiles: (1) La constancia del reloj molecular para una proteína dada (a continuación) y (2) la paradoja de los niveles de variación genética (Lewontin 1974) (véase sección 1.2.2).

### RELOJ MOLECULAR vs TEORÍA NEUTRALISTA

La constancia del reloj molecular, también conocido como el efecto del tiempo de generación (*generation-time effect*), es sorprendente si se considera que la unidad natural de medida del tiempo en la teoría neutralista es la generación (no los años). Como consecuencia, bajo la teoría neutralista especies con generaciones cortas deberían evolucionar más rápido (en tiempo real, años) que aquellas especies con generaciones más largas. No obstante, se observó que los relojes moleculares basados en proteínas no cumplen esta predicción, es decir, el reloj molecular proteico parece bastante constante entre especies cuando no debería serlo (Zuckerkandl y Pauling 1965; Wilson *et al.* 1977). Ohta y Kimura (1971) sugirieron una solución a este problema; esta solución es la teoría casi neutra de la evolución molecular. Sugirieron que la *DFE* debería ser continua (véase figura 1.3C), en lugar de estar compuesta por tres categorías: deletéreas, neutras y beneficiosas (aunque esta última categoría es mucho menos importante cuantitativamente que las otras dos). Razonaron que mutaciones deletéreas con coeficientes de selección menores a  $1/N_e$ , el reciproco de  $N_e$ , están sujetas a la deriva genética y se comportan como neutras (Kimura 1968) (véase figura 1.2), en cambio aquellas mutaciones con coeficientes de selección mayores a  $1/N_e$  se comportan como deletéreas. Especies con tamaños efectivos mayores tienen por lo tanto una menor proporción de mutaciones efectivamente neutras, o casi neutras, que especies con  $N_e$  menores. Si

además estas tienen tasas de mutación mayores por unidad de tiempo real (año) porque tienen tiempos de generación más cortos, entonces la tasa de evolución resultante será constante por año entre especies. De hecho, hay evidencias que sugieren que esto último es cierto, es decir, que la tasa de mutación por año es mayor para especies con tiempos de generación cortos (Li *et al.* 1987; Mooers *et al.* 1994; Bromham *et al.* 1996). En otras palabras, especies con  $N_e$  mayores tienen una menor proporción de mutaciones efectivamente neutras ( $f$ ) y una mayor tasa de mutación por año ( $\mu$ ) (porque tienen generaciones cortas), como la tasa de substitución neutra depende del producto de estas dos variables,  $K = f \mu$  (Kimura 1983), ambas variables se cancelan la una a la otra y la  $K$  resultante aparece constante por año entre especies. A partir de entonces Ohta pasó a defender la teoría casi neutra en una serie de trabajos (Ohta 1972b; 1973; 1976; 1977; 1992).

La teoría casi neutra incorpora dos nuevas categorías de mutaciones respecto a la teoría neutralista. Por un lado, las mutaciones casi neutras o efectivamente neutras (*nearly neutral* o *effectively neutral*). Estas son mutaciones tanto deletéreas como beneficiosas, y puramente neutras, con coeficientes de selección comprendidos entre los límites impuestos por el reciproco de  $N_e$  ( $-1 < N_e s < 1$ ). Por otro lado, las mutaciones ligeramente deletéreas (*slightly* o *weakly deleterious*) con coeficientes de selección entre la frontera de lo fuertemente deletéreo ( $N_e s < -10$ ) y de lo efectivamente neutro ( $-10 < N_e s < -1$ ) (Ohta y Kimura 1971; Ohta 1972a, 1995; Li 1987). Estas son mutaciones que contribuyen al polimorfismo pero no a la divergencia (o muy poco). Las evidencias a favor de la presencia de las mutaciones casi o efectivamente neutras ( $-1 < N_e s < 1$ ) están basadas en la correlación positiva entre el constreñimiento proteico y el censo efectivo,  $N_e$  (figura 1.4). Las evidencias a favor de la presencia de las mutaciones ligeramente deletéreas ( $-10 < N_e s < -1$ ) están basadas en los patrones de polimorfismo no-sinónimo respecto al sinónimo.

## MUTACIONES NO-SINÓNIMAS EFECTIVAMENTE NEUTRAS

El constreñimiento ( $C$ ) se calcula como 1 menos la razón entre la tasa de substitución no-sinónica (o aminoacídica) ( $d_H$ ) y la tasa de substitución sinónica ( $d_S$ ),  $C = 1 - d_H/d_S$ . En un modelo donde las sustituciones sinónimas son neutras y las sustituciones no-sinónimas pueden ser neutras o deletéreas, el constreñimiento es la proporción de mutaciones puntuales de cambio de aminoácido que son deletéreas y eliminadas por la selección natural. Ohta 1972a encontró por primera vez una correlación entre el constreñimiento proteico y un potencial indicador o *proxy* de  $N_e$ , el tiempo de generación (figura 1.4). Años más tarde se observó efectivamente que el tiempo de generación está correlacionado negativamente con  $N_e$  (Chao y Carr 1993). Por lo tanto, este trabajo inicial de Ohta fue el primero en encontrar que el constreñimiento está correlacionado positivamente con  $N_e$  en un abanico amplio de especies de mamíferos y moscas del género *Drosophila*.

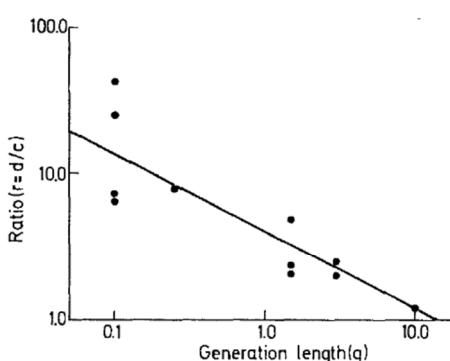


Fig. 1. The relationship between the generation time ( $g$ ) and the ratio ( $r$ ) of DNA divergence to cistron divergence both in logarithmic scale. From the regression line we get  $r\sqrt{g} \approx 4.5$ . The value of  $r$  represents approximately the reciprocal of the proportion of accepted mutations in cistrons

**FIGURA 1.4** Primera evidencia a favor de la teoría casi neutra de la evolución molecular (Ohta 1972a). Se muestra el primer resultado que sugiere que el número de nuevas mutaciones no-sinónimas que son casi neutras ( $-1 < N_S < 1$ ) dependen de  $N_e$ . En el eje de la  $x$  se representa el logaritmo del tiempo de generación, el cual está negativamente correlacionado con  $N_e$ , por lo tanto, en los valores cercanos a 0 tenemos las especies con mayor  $N_e$  y cerca del 10 las especies con menor  $N_e$ . En el eje de la  $y$  se representa el logaritmo de la razón entre la tasa de sustitución en el ADN no-codificador (*DNA divergence*) y la tasa de substitución no-sinónica (*cistrons divergence*). Los valores cercanos a 1 están menos constreñidos que los valores cercanos a 100. [Figura tomada de Ohta (1972a)].

Una serie de trabajos posteriores corroboraron los resultados de Ohta en un número mayor de especies (Li *et al.* 1987; Ohta 1995; Keightley y Eyre-Walker 2000), e incluso encontraron que para especies de *Drosophila* (Ohta 1976; 1993) y pájaros (Johnson y Seger 2001) que habitan en islas el constreñimiento es menor que en las especies continentales. En Eyre-Walker *et al.* (2002) se comparó los niveles de constreñimiento entre parejas de especies cercanas, pero con  $N_e$  actuales muy diferentes, observaron que el constreñimiento era menor en aquellas especies con un  $N_e$  actual más bajo.

### MUTACIONES NO-SINÓNIMAS LIGERAMENTE DELETÉREAS

La fuente de evidencias a favor de la existencia de mutaciones ligeramente deletéreas proviene de la variación genética intra-poblacional. Se ha observado un exceso de la razón entre el polimorfismo no-sinónimo ( $p_N$ ) y el polimorfismo sinónimo ( $p_S$ ) respecto a la razón entre la divergencia no-sinónima ( $d_N$ ) y la divergencia sinónima ( $d_S$ ), es decir,  $p_N/p_S \gg d_N/d_S$  en genomas mitocondriales (Rand y Kann 1996; Nachman 1998), en genes nucleares de *Arabidopsis thaliana* (Weinreich y Rand 2000), de humanos (Bustamante *et al.* 2005), de *Drosophila* (Presgraves 2005; Welch 2006) y en bacterias entéricas (Charlesworth y Eyre-Walker 2006). Este patrón es consistente con la segregación de mutaciones no-sinónimas ligeramente deletéreas, las cuales contribuyen al polimorfismo pero rara vez se fijan (Kimura 1983, p. 44). Además el espectro de frecuencias (o *Site Frequency Spectrum, SFS*) de las mutaciones no-sinónimas está inflado en variantes a baja frecuencia en comparación con las mutaciones neutras en el genoma mitocondrial de diversas especies (Nielsen y Weinreich 1999), en el genoma nuclear de humanos (Cargill *et al.* 1999; Sunyaev *et al.* 2000; Altshuler *et al.* 2010; Li *et al.* 2010), *Drosophila* (Akashi 1999; Fay *et al.* 2002; Begun *et al.* 2007; Langley *et al.* 2012; Mackay *et al.* 2012; Pool *et al.* 2012) y bacterias (Hughes 2005; Charlesworth y Eyre-Walker 2006). Este hecho es también indicativo de la presencia de alelos ligeramente deletéreos segregando en las poblaciones.

Métodos estadísticos avanzados, capaces de estimar la *DFE* indirectamente a partir de alineamientos de ADN comparando el SFS de mutaciones neutras y selectivas (Eyre-Walker *et al.* 2006; Keightley y Eyre-Walker 2007; Boyko *et al.* 2008) (véase sección 1.3.3), han resuelto que alrededor de un 30-40% de las nuevas mutaciones no-sinónimas son efectivamente neutras ( $-1 < N_e s < 1$ ) en humanos y sólo un 6% lo son en *Drosophila* (Eyre-Walker y Keightley 2009), mientras que entre un 6-18% son ligeramente deletéreas ( $-10 < N_e s < -1$ ) en humanos y alrededor de un 7% lo son en *Drosophila*. Los ratones muestran unos valores intermedios, con un 15% de las nuevas mutaciones efectivamente neutras ( $-1 < N_e s < 1$ ) y un 11% ligeramente deletéreas ( $-10 < N_e s < -1$ ) (Kousathanas *et al.* 2011). Aplicando estos métodos a distintas especies de plantas Gossmann *et al.* (2010) cuantificó que la fracción de nuevas mutaciones no-sinónimas ligeramente deletéreas es sorprendentemente constante, alrededor del 5-15%, mientras que la fracción de mutaciones efectivamente neutras parecen ser más variables y depender de los eventos de domesticación.

## MUTACIONES NO-CODIFICADORAS

¿Qué hay del ADN no-codificador? ¿Qué sabemos de la *DFE* y de la presencia de mutaciones ligeramente deletéreas y efectivamente neutras predichas por Ohta en este caso? Hasta hace relativamente poco la mayoría de los biólogos evolutivos consideraban que el ADN no-codificador evolucionaba de forma neutra. No obstante, esta visión es muy distinta en la actualidad. La mayoría de la información sobre la *DFE* para mutaciones que ocurren fuera de la secuencia codificadora provienen de las estimas del constreñimiento selectivo (es decir, comparando niveles de divergencia respecto sitios neutros, véase más arriba), aunque recientemente los mismos métodos de estimación de la *DFE* a partir del SFS que se han aplicado a las secuencias codificadoras están siendo aplicados al genoma no-codificador. En levadura y nemátodos alrededor de un 10-20% del genoma no codificador está sometido a

selección purificadora (Shabalina y Kondrashov 1999; Cliften *et al.* 2003). En *D. melanogaster* un ~50% de los sitios no-codificadores parecen estar sometidos a la acción de la selección natural negativa ( $N_e s < -1$ ) (Bergman y Kreitman 2001; Andolfatto 2005; Halligan y Keightley 2006). Un estudio sobre la DFE en secuencias intrónicas de *Drosophila* basadas en el SFS apuntó que un 70% de las nuevas mutaciones puntuales son efectivamente neutras ( $-1 < N_e s < 1$ ), un 23% ligeramente deletéreas ( $-10 < N_e s < -1$ ) y un 7% fuertemente deletéreas ( $N_e s < -10$ ) (Eyre-Walker y Keightley 2009). Otro estudio en *Drosophila* donde se analizaron exclusivamente secuencias no-codificadoras muy conservadas (*highly conserved non-coding sequences, CNS*) indicó que alrededor de un 85% de las nuevas mutaciones puntuales que ocurren en estos sitios son muy deletéreas ( $N_e s < -10$ ) (Casillas *et al.* 2007). En mamíferos las estimas son bastante inferiores, alrededor del 5-7% del ADN no-codificador parece estar constreñido (Mouse Genome Sequencing Consortium 2002; Dermitzakis *et al.* 2005; Gaffney y Keightley 2006; Davydov *et al.* 2010; Lindblad-Toh *et al.* 2011). En ratones, estudios centrados en las regiones no-codificadoras 500 pb aguas arriba y aguas abajo del primer y último codón, respectivamente, señalan que un 71% de las mutaciones son efectivamente neutras ( $-1 < N_e s < 1$ ), un 9% ligeramente deletéreas ( $-10 < N_e s < -1$ ) y un 20% fuertemente deletéreas ( $N_e s < -10$ ) (Kousathanas *et al.* 2011). En humanos, sin embargo, estudios centrados en las regiones no-codificadoras 500 pb aguas arriba y aguas abajo del primer y último codón, respectivamente, no son capaces de estimar mutaciones seleccionadas (con  $N_e s < -1$ ) (Eyre-Walker y Keightley 2009). Esto encajaría con el hecho que el nivel de constreñimiento en estas regiones para homínidos es mucho menor que para roedores (Keightley *et al.* 2005a; Bush y Lahn 2005). Esto es seguramente así porque la selección ha sido menos eficiente en los homínidos debido a su menor  $N_e$  a largo plazo. Es más, dentro de bloques conservados de ADN no-codificador entre homínidos y roedores, los homínidos tienen niveles menores de conservación que los roedores (Keightley *et al.* 2005b; Kryukov 2005).

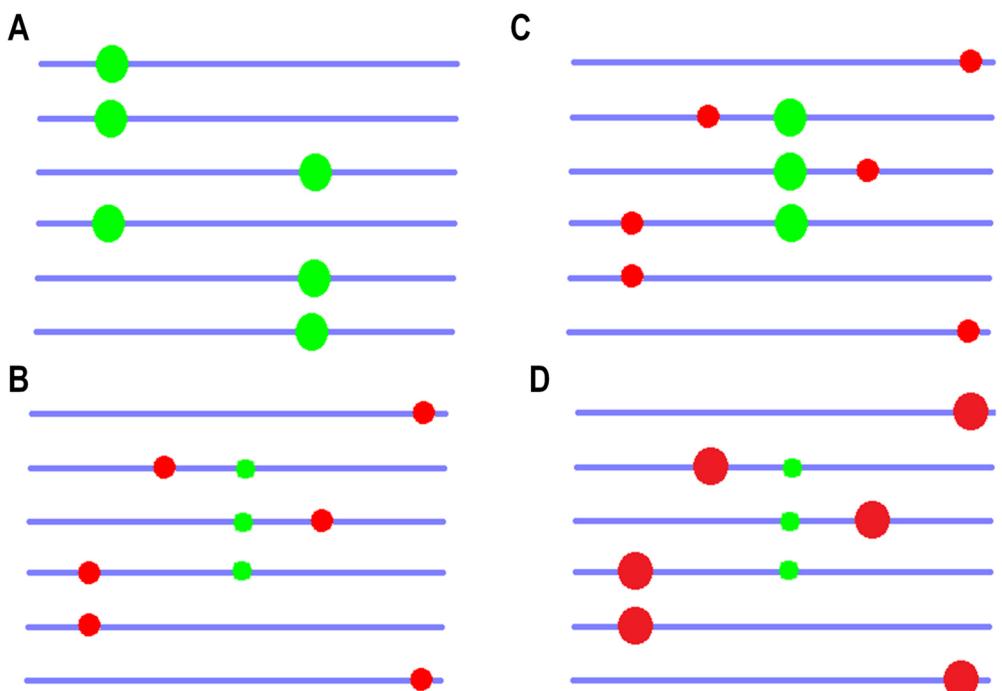
A pesar de todos estos avances, la teoría casi neutra tiene todavía mucho camino que recorrer. La mayoría de las estimas de la *DFE* basadas en el SFS se han centrado en secuencias codificadoras de humanos (Eyre-Walker y Keightley 2009), *Drosophila* (Eyre-Walker y Keightley 2009), ratones (Kousathanas *et al.* 2011) y algunas especies de plantas (Gossmann *et al.* 2010) para mutaciones puntuales. Sin embargo, es muy probable que otro tipo de mutaciones, como las inserciones de elementos transponibles o las inversiones, tengan *DFEs* completamente distintas (Eyre-Walker y Keightley 2007). De hecho, el lugar donde ocurren las mutaciones puntuales tiene un enorme efecto sobre sus efectos sobre la eficacia biológica; la proporción de mutaciones puntuales casi neutras en el ADN no-codificador parece ser mayor que en el ADN codificador en todas las especies estudiadas. Por si fuera poco, los métodos utilizados para inferir la *DFE* a partir del SFS (Eyre-Walker *et al.* 2006; Keightley y Eyre-Walker 2007; Boyko *et al.* 2008) asumen que esta sigue una distribución continua, cuando en realidad podría no hacerlo tal y como discuten Kousathanas y Keightley (2013). Si la *DFE* no sigue una distribución continua muchos trabajos y conclusiones al respecto de la *DFE* deberán revisarse. Es decir, lo poco que sabemos de los detalles de la *DFE* podría no ser cierto. Pero el caso es “peor” si cabe para la inferencia de la *DFE* de nuevas mutaciones beneficiosas, ya que éstas son muy escasas y permanecen poco tiempo segregando en las poblaciones (véase figura 1.2B). Un estudio basado en el contraste del SFS de sitios sinónimos y no-sinónimos (Scheneider *et al.* 2011) ha estimado que en *Drosophila* la mayoría de las nuevas mutaciones no-sinónimas adaptativas son ligeramente adaptativas ( $N_{eS} \sim 2.5-5$ ) y que alrededor de un 1-2% de las nuevas mutaciones no-sinónimas son beneficiosas. Todas estas incógnitas sobre la forma de la *DFE* para distintos tipos de mutaciones en distintas especies indican que hay una necesidad creciente de revisión y mejora de nuestro conocimiento sobre la teoría casi neutra y su alcance a lo largo del árbol de la vida.

En conclusión, numerosas evidencias sugieren que la teoría casi neutra de la evolución molecular de Ohta (Ohta 1972b; 1973; 1976; 1977; 1992) es el mejor modelo del que disponemos actualmente para explicar los patrones generales de variación genética entre y dentro de poblaciones. Esto no quiere decir que no haya determinados *loci* bajo selección equilibradora o selección positiva recurrente, simplemente indica que la mayoría de alelos no se encuentran bajo selección natural directa y los que se encuentran seleccionados lo están débilmente.

### 1.1.3 EFECTO HILL-ROBERTSON

Los genomas de la mayoría de organismos contienen miles de genes, no obstante, el grueso de la teoría evolutiva molecular considera que los sitios (o *loci*) segregan independientemente. Asumir no sólo que los genes evolucionan de manera independiente entre sí, sino que las mutaciones que ocurren dentro de un mismo gen también lo hacen es un supuesto que raramente se cumple (a excepción de algunos virus con unas tasas de recombinación muy elevadas). Ahora bien, ¿cómo afecta la falta de independencia en la segregación a las predicciones obtenidas anteriormente con el modelo de Wright-Fisher? Idealmente para responder esta cuestión deberíamos modelar el genoma completo. Desafortunadamente, no hay un tratamiento analítico general que describa la dinámica de la interferencia entre múltiples alelos bajo selección y deriva genética, para distintos coeficientes de selección y niveles de recombinación. Sin embargo, podemos extraer mucha información simplemente modelando dos *loci* ligados. Esto es lo que hicieron Hill y Robertson en un estudio pionero en 1966. Descubrieron un fenómeno que pasó a denominarse a partir de Felsenstein (1974) efecto Hill-Robertson. Gracias a los avances en computación este efecto ha podido ser estudiado en múltiples *loci* mediante simulaciones *forward in time* asumiendo el modelo de Wright-Fisher (sección 1.1.1) (Comeron *et al.* 1999; Tachida 2000, McVean y Charlesworth 2000; Comeron y Kreitman 2000; 2002; Messer y Petrov 2013). En todos estos trabajos

siempre entran en juego tres parámetros: (1) el número de sitios bajo selección ( $L$ ), (2) la intensidad de la selección ( $N_s s$ ) y (3) la tasa de recombinación ( $\rho$ ). El efecto Hill-Robertson puede definirse de manera genérica como la interferencia que se da en una población finita entre dos o más alelos seleccionados (sin necesidad de epistasis) que están en desequilibrio de ligamiento. Si el tamaño de población es infinito y/o la selección ocurre sobre variantes preexistentes, la misma mutación seleccionada se encontrará en distintos fondos (*backgrounds*) genéticos reduciendo el desequilibrio de ligamiento y la posibilidad de interferencia entre mutaciones (Maynard Smith 1968; Felsenstein 1974; Roze y Barton 2006).



**FIGURA 1.5** Escenarios efecto Hill-Robertson. Los cromosomas están representados como barras horizontales. Los círculos de color rojo simbolizan mutaciones deletéreas y los círculos verdes mutaciones adaptativas; los círculos pequeños son mutaciones ligeramente seleccionadas y los grandes mutaciones fuertemente seleccionadas. (A) Corresponde al escenario 1.1, interferencia simétrica entre dos o más alelos seleccionados positivamente; (B) corresponde al escenario 1.2, interferencia simétrica entre un alelo seleccionado positivamente y otro alelo seleccionado negativamente; mientras (C) y (D) corresponden al escenario 2.1, interferencia asimétrica con dominio selección positiva y 2.2 interferencia asimétrica con dominio selección negativa, respectivamente. Estos escenarios están descritos en detalle en el texto principal.

La interferencia de Hill-Robertson puede darse en distintos escenarios (figura 1.5) entre dos o más *loci*:

**Escenario 1:** Interferencia simétrica, los alelos tienen coeficientes de selección equivalentes.

**Escenario 1.1:** Interferencia entre dos o más alelos seleccionados positivamente. Tradicionalmente se han asumido mutaciones fuertemente adaptativas ( $N_e s > 10$ ) aunque si la recombinación es muy baja o nula podría aplicar también para mutaciones débilmente seleccionadas ( $N_e s \sim 1-10$ ). En organismos asexuales donde la recombinación está ausente o es muy poco común también se le denomina interferencia clonal (Fisher 1930; Crow y Kimura 1965). Considérese una población haploide con dos *loci* completamente ligados. Cada *locus* tiene un alelo benéfico que aumenta la eficacia biológica una fracción  $s$ . Ambos *loci* tienen inicialmente una copia del alelo benéfico. Por lo tanto, en  $1-1/N$  ocasiones los dos alelos se encontrarán en cromosomas diferentes (o en desequilibrio de ligamiento negativo o repulsión) y en  $1/N$  ocasiones se encontrarán en el mismo cromosoma (o en desequilibrio de ligamiento positivo). En el primer caso el desequilibrio de ligamiento negativo disminuirá la probabilidad de fijación de los alelos y con ello la tasa de evolución adaptativa. En el segundo caso, los dos alelos benéficos se ayudarán el uno al otro y la probabilidad de fijación aumentará para ambos. Podríamos pensar que ambos efectos se cancelan el uno al otro y al final del día la tasa de adaptación cuando las mutaciones están ligadas es la misma que cuando las mutaciones son independientes. Sin embargo, la deriva genética crea desequilibrio de ligamiento al azar entre dos *loci* y el desequilibrio de ligamiento negativo vence (porque dura más tiempo) al desequilibrio de ligamiento positivo (Roze y Barton 2006). El resultado, la probabilidad global de fijación de nuevas mutaciones adaptativas disminuye (Hill y Robertson 1966; Felsenstein 1974). Esto sucede también cuando la segunda mutación

adaptativa ocurre poco tiempo después de la primera (Crow y Kimura 1965).

**Escenario 1.2:** Interferencia entre mutaciones beneficiosas y deletéreas débilmente seleccionadas ( $N_e s \sim |1-10|$ ). Este escenario ha sido denominado por algunos autores como *Interference Selection* (IS) (por Comeron y Kreitman 2002) o *weak selection Hill-Robertson* (wsHR) (por McVean y Charlesworth 2000). La probabilidad de fijación de las mutaciones deletéreas aumenta y el de las beneficiosas disminuye.

**Escenario 2:** Interferencia asimétrica, uno de los alelos tiene un coeficiente de selección mayor al otro. Siempre se asume que uno de los dos alelos está débilmente seleccionado (esto es  $N_e s \sim |1-10|$ ).

**Escenario 2.1:** El alelo seleccionado positivamente tiene un coeficiente de selección mayor al seleccionado negativamente. En este caso la probabilidad de fijación del alelo deletéreo aumenta y la del alelo beneficioso disminuye ligeramente.

**Escenario 2.2:** El alelo seleccionado negativamente tiene un coeficiente de selección mayor al seleccionado positivamente. En este caso la probabilidad de fijación del alelo adaptativo disminuye y la del alelo deletéreo aumenta ligeramente.

**Escenario 2.3:** Alelos deletéreos seleccionados fuerte y débilmente. En este caso se produce una reducción global del  $N_e$  y la probabilidad de fijación de todas las mutaciones aumenta. Sin embargo, las mutaciones más débilmente seleccionadas contribuyen más a la divergencia. El hecho que no se haya estudiado la interferencia entre mutaciones deletéreas fuertemente seleccionadas es debido a que estas permanecen muy poco tiempo segregando en la población y es muy poco probable que interfieran unas con otras.

A pesar de esta variedad de escenarios ¿cómo se cuantifica la intensidad del efecto Hill-Robertson en la práctica? Trabajos teóricos, por ejemplo, han estimado la razón entre la probabilidad de fijación del alelo benéfico con interferencia respecto a la probabilidad de fijación del alelo benéfico sin interferencia para distintas combinaciones de parámetros (Barton 1995). Una razón igual a uno equivale a la ausencia de evidencias del efecto Hill-Robertson, una razón menor a uno nos indica la acción del efecto Hill-Robertson. En otras palabras, el efecto Hill-Robertson puede cuantificarse como la fracción de substituciones adaptativas que dejan de fijarse (Barton 1995). Sin embargo, en la práctica este estadístico nunca había sido estimado sobre un genoma. Precisamente, la estima de dicho estadístico sobre el genoma de *D. melanogaster* es uno de los objetivos de esta tesis (secciones 1.5 y 3.3.4). Existen otros estadísticos para cuantificar la intensidad del efecto Hill-Robertson basados en los niveles de polimorfismo, el sesgo en el espectro de frecuencias y el desequilibrio de ligamiento, los cuales han sido aplicados sobre genes de *D. melanogaster* para demostrar que la interferencia de Hill-Robertson es más intensa en la región central de genes sin intrones (Comeron y Kreitman 2002).

Poco se sabe de la importancia relativa de cada uno de los escenarios (representados en la figura 1.5) en el mundo real, pues la prevalencia de uno u otro depende de la DFE. Cabe destacar que las mutaciones efectivamente neutras y ligeramente seleccionadas ( $-10 < N_{es} < 10$ ), ya sean benéficas o perjudiciales, permanecen más tiempo segregando en las poblaciones a frecuencias más altas y, por lo tanto, son las más susceptibles a sufrir el efecto Hill-Robertson (McVean y Charlesworth 2000; Comeron y Kreitman 2002; Comeron *et al.* 2008). Además, cuando los coeficientes de selección de los alelos implicados en la interferencia son asimétricos, el alelo más fuertemente seleccionado (con  $|N_{es}| > 10$ ) parece no sufrir tanto las consecuencias de la iHR (McVean y Charlesworth 2000; Johnson y Barton 2002; Charlesworth 2012a, no obstante, véase Messer y Petrov 2013). En cualquier caso, todos los trabajos apuntan a que el efecto Hill-Robertson es siempre más intenso cuanto menor es la

distancia genética entre los alelos seleccionados, o mayor es la densidad de sitios seleccionados por unidad de distancia genética (Li 1987; McVean y Charlesworth 2000; Comeron y Kreitman, 2002; Comeron *et al.* 2008).

### EFFECTO HILL-ROBERTSON Y UNIDADES GENÓMICAS ASOCIADAS AL $N_e$

Retornando a la pregunta planteada previamente: ¿Cómo afecta la falta de independencia en la segregación a las predicciones obtenidas anteriormente en el modelo de Wright-Fisher? Podemos responder que el hecho que las mutaciones no segreguen de manera independiente genera desviaciones en las predicciones del modelo de Wright-Fisher como mínimo en lo que respecta a las probabilidades de fijación de nuevas mutaciones seleccionadas. La falta de segregación independiente entre sitios a lo largo de los cromosomas conduce a un aumento de la probabilidad de fijación de nuevas mutaciones deletéreas y a una disminución de la probabilidad de fijación de nuevas mutaciones beneficiosas. No obstante, estas desviaciones respecto a lo esperado bajo segregación independiente pueden corregirse contemplando un censo efectivo,  $N_e$ , variable a lo largo del genoma el cual es el producto de la interacción entre: (1) el número de sitios bajo selección ( $L$ ), (2) la intensidad de la selección ( $N_e s$ ) y (3) la tasa de recombinación ( $\rho$ ). Así pues, no tan sólo hay un  $N_e$  a nivel de población, o especie específico, como el que contempló Wright (1931, 1938b) (sección 1.1.1) originalmente, sino también un  $N_e$  a escala subcromosómica, gen específica e incluso mutación específica. El efecto Hill-Robertson se ha interpretado tradicionalmente como un aumento del efecto de la deriva genética localmente (Robertson 1961; Birky y Walsh 1988), que conlleva a una reducción del  $N_e$  y de la eficiencia de la selección sobre los alelos implicados (Robertson 1961; Hill y Robertson 1966; Felsenstein 1974; Kliman y Hey 1993).

Más recientemente Messer y Petrov (2013) simularon cromosomas con características propias de humanos, encontraron que en un escenario con arrastres

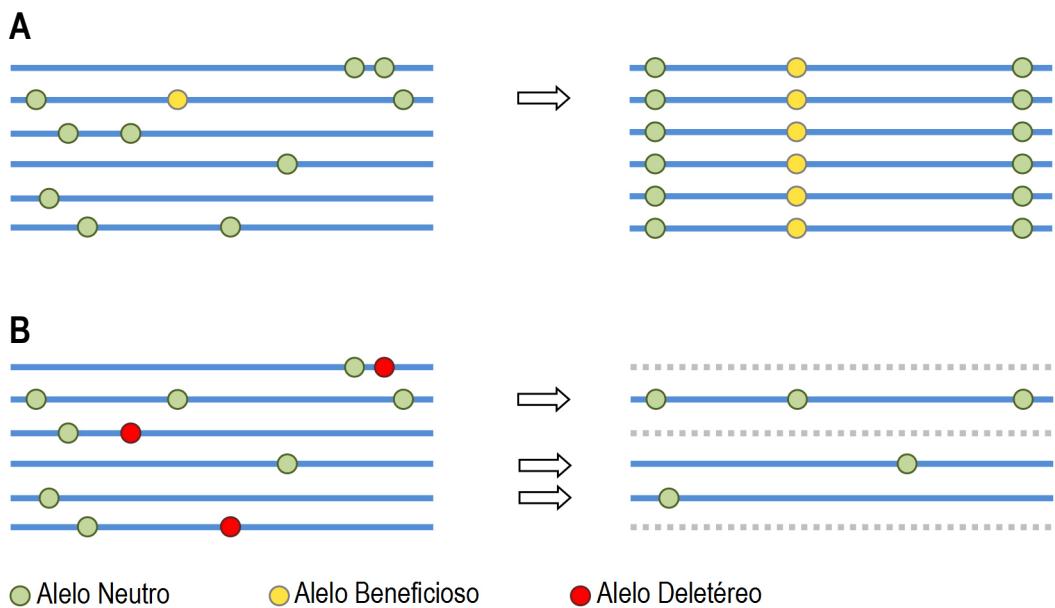
positivos y negativos recurrentes la reducción en  $N_e$  es mayor para mutaciones fuertemente deletéreas que para las débilmente deletéreas. Es decir, no encontraron que un único valor de  $N_e$  fuera aplicable a todos los valores de coeficientes de selección estudiados. Además, el  $N_e$  observado en algunos casos era hasta un orden de magnitud menor al  $N_e$  simulado (consúltese figura 1C Messer y Petrov 2013). Cabe la posibilidad que genomas más densos funcionalmente como el de *Drosophila*, donde el ~50% de los sitios parecen estar sometidos a selección (Bergman y Kreitman 2001; Andolfatto 2005; Halligan y Keightley 2006) esta variación de  $N_e$  a lo largo del genoma pueda ser mayor que en el caso humano. No obstante, la tasa de recombinación (y el decaimiento del desequilibrio de ligamiento) por sitio en *Drosophila* es mucho mayor que en humanos. En definitiva, se requieren más simulaciones para conocer como la interacción entre todas estas variables acaba afectando el  $N_e$  localmente en genomas como el de *Drosophila* o humanos.

En conclusión, el efecto Hill-Robertson puede generar fluctuaciones considerables del  $N_e$  (y la eficiencia de la selección) localmente, pero la reducción local del  $N_e$  no es la misma para todas las mutaciones, depende de su coeficiente de selección. Dentro de un mismo cromosoma podemos encontrarnos regiones con un  $N_e$  entre 10-100 veces mayor (o menor) a otra.

### EFFECTO HILL-ROBERTSON vs SELECCIÓN LIGADA

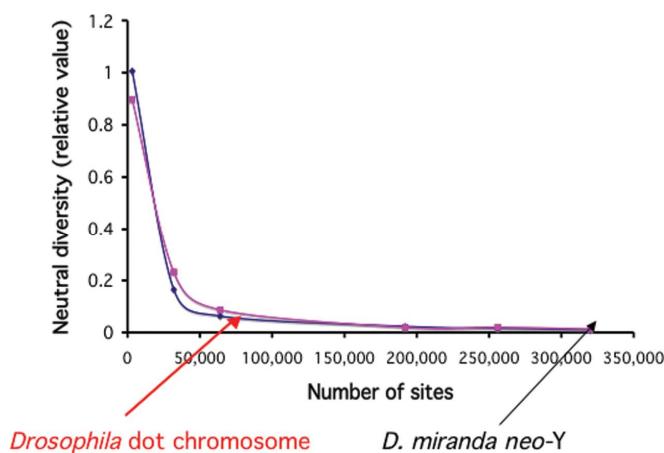
El efecto Hill-Robertson se encuentra íntimamente relacionado con otro concepto, la selección ligada. En algunos artículos ambos términos tienden a confundirse. Sin embargo, el efecto Hill-Robertson trata sobre el efecto que variantes seleccionadas tienen sobre otras variantes seleccionadas ligadas, mientras que la selección ligada se refiere al efecto que las variantes seleccionadas tienen sobre otras variantes neutras ligadas. Es un hecho que los niveles de variación neutra están positivamente correlacionados con la tasa de recombinación en multitud de especies (véase sección

1.2.1 para más detalles). Dos modelos teóricos se han propuesto para explicar dicha correlación: (1) el efecto autoestopista (HH, *Hitchhiking*, Maynard Smith y Haigh 1974), donde las variantes neutras están ligadas a un alelo beneficioso fuertemente seleccionado, y (2) la selección de fondo (BGS, *Background Selection*, Charlesworth *et al.* 1993), donde las variantes neutras están ligadas a un alelo efectivamente deletéreo (figura 1.6).



**FIGURA 1.6** Efectos de los arrastres selectivos sobre la variación neutra. (A) El arrastre de alelos neutros ligados a un alelo beneficioso (*Hitchhiking*, HH) reduce la variación neutra; la recombinación rompe el haplotipo portador del alelo beneficioso y con el tiempo la mutación re establece los niveles de variación neutra. La razón entre la tasa de recombinación entre el alelo neutro y el alelo beneficioso ( $r$ ) y el coeficiente de selección adaptativo ( $s$ ),  $r/s$ , es muy importante para determinar la reducción en la variabilidad neutra (Maynard Smith y Haigh 1974; Barton 2000). Consultese sección 1.3.1 para ver otros ejemplos de arrastres positivos: *soft sweeps* y *partial sweeps*. (B) La selección de fondo (*Background selection*, BGS) o arrastre selectivo negativo ocurre cuando alelos neutros se encuentran en el mismo cromosoma que alelos deletéreos destinados a la extinción (Charlesworth *et al.* 1993). Los niveles de variación neutra se reducen menos que bajo el arrastre selectivo positivo (Charlesworth 2012a; Comeron 2014). [Figura adaptada de Ràmia (2015)].

Ambos modelos predicen una correlación positiva entre los niveles de polimorfismo y la recombinación. Por lo tanto, no está claro qué modelo, HH o BGS, domina los patrones de variación neutra (véase sección 1.2.1). Por otra parte, las mutaciones débilmente seleccionadas (positiva o negativamente) se cree alteran poco los niveles de variación de mutaciones neutras ligadas (Golding 1997; Neuhauser y Krone 1997; Przeworski *et al.* 1999). No obstante, Comeron y Kreitman (2002) y Charlesworth (2013a) encontraron que cuando múltiples mutaciones de este tipo se encuentran ligadas a mutaciones neutras los niveles de variación en estas últimas también pueden verse significativamente reducidos.



**FIGURA 1.7** Efecto de la BGS sobre la diversidad neutra en un cromosoma sin entrecruzamiento. La tasa de mutación es la típica de *Drosophila* ( $\mu \sim 3.5 \times 10^{-9}$  [Keightley *et al.* 2009]). El eje de la x representa el número de sitios codificadores en el cromosoma; son tripletes o codones, donde los dos primeros sitios están sometidos a selección y el tercero es neutro. En el eje de la y se representa la razón entre la  $n$  observada y la  $n$  esperada con libre recombinación. La curva azul es el resultado de la simulación sin recombinación (ni entrecruzamiento ni conversión génica), la curva lila representa el resultado de la simulación con la tasa de conversión génica de *Drosophila* (tracto de conversión promedio ~500 pb, Comeron *et al.* 2012). Las flechas indican los niveles de  $n$  observados para el cromosoma *dot* y el *neo-Y* de *D. miranda* y su relación con el número aproximado de sitios codificadores en estos cromosomas. [Figura tomada de Charlesworth (2013a)].

Charlesworth (2013a) simuló el efecto de la BGS sobre la variación neutra en ausencia total de recombinación (como ocurre en muchos genomas bacterianos y cromosomas sexuales de muchas especies) y relacionó el número de sitios bajo selección purificadora débil ( $-10 < N_e S < -1$ ) con la reducción de la heterocigosidad nucleotídica

de mutaciones neutras,  $\pi$  (Tajima 1983). Charlesworth (2013a) mostró como a partir de ~50.000 sitios codificadores  $\pi$  se reduce hasta unas 100 respecto al valor esperado bajo libre recombinación (figura 1.7). Se puede observar como el modelo de BGS predice perfectamente los niveles de variación neutra observados en el cromosoma puntual (*dot*) y *neo-Y* de *D. miranda*. Finalmente, a continuación (cuadro 3), se explica cómo afecta la selección en sitios ligados a la probabilidad de fijación de mutaciones neutras.

### CUADRO 3: SELECCIÓN EN SITIOS LIGADOS Y LA PROBABILIDAD DE FIJACIÓN DE MUTACIONES NEUTRAS

El grueso de la teoría evolutiva asume, entre otros supuestos que posibilitan el tratamiento matemático, que los sitios evolucionan de manera independiente. Ya se conocía que cuando los sitios segregan de manera independiente la tasa de substitución neutra ( $K$ ) es igual al producto de la tasa de mutación por generación ( $\mu$ ) y la proporción de mutaciones que son neutras ( $f$ ),  $K = f \mu$  (Kimura 1983). Watterson (1982) demostró que la ausencia completa de recombinación no afecta a la tasa de substitución neutra cuando todas las mutaciones son neutras. Birky y Walsh (1988) demostraron además que esto es así incluso en presencia de sitios ligados bajo selección. Mostraron analíticamente (cuando la recombinación es igual a 0) y mediante simulaciones (para valores de recombinación  $> 0$ ) que la probabilidad de fijación de mutaciones neutras ligadas a otras mutaciones deletéreas o beneficiosas no se ve afectada por la selección que ocurre sobre sitios adyacentes.

La intuición de por qué esto es así es sencilla. La selección sobre sitios ligados reduce  $N_e$  localmente, sin embargo, la tasa de substitución no depende de  $N_e$ , sino de la frecuencia inicial (Kimura 1983). Si bien es cierto que un arrastre selectivo positivo podrá aumentar la frecuencia de muchas variantes neutras localizadas en el mismo cromosoma, y con ello su probabilidad de fijación, también es cierto que muchas variantes neutras a alta frecuencia que estaban a punto de fijarse dejarán de hacerlo si la mutación adaptativa ha ocurrido sobre otro cromosoma. Es decir, ambos procesos se cancelan el uno al otro. Para el escenario con selección de fondo la idea es que, aunque muchas variantes neutras se purgarán junto con la variante deletérea ligada, y con ello aumentara la varianza en el número de copias de los distintos alelos neutros, este efecto no afectará a la probabilidad de fijación de las variantes neutras, la cual sólo depende de la frecuencia inicial (Kimura 1983).

Un aspecto importante de la variación del  $N_e$  a lo largo del genoma y la tasa de substitución neutra, es que los niveles de divergencia neutra entre especies muy cercanas como humanos y chimpancés o *D. melanogaster* y *D. simulans*, sí pueden verse afectados por la variación en el  $N_e$  (Reed *et al.* 2005). El polimorfismo ancestral contribuye a una fracción importante del total de fijaciones neutras que encontramos en un linaje a lo largo de  $10N_e$  generaciones (Charlesworth *et al.* 2005). Esto representa 2.5 millones de años en el caso de los humanos. Atribuir polimorfismo ancestral neutro al contaje de sustituciones puede inflar nuestras estimas de la divergencia respecto a las estimas que esperaríamos a través de la fijación de nuevas mutaciones. En aquellas regiones del genoma de la población ancestral donde el  $N_e$  era pequeño, debido a la selección ligada, habrá menos polimorfismo ancestral y menores estimas de divergencia en las especies actuales. Esto generará en la actualidad una correlación positiva entre polimorfismo y divergencia neutra. Es importante tener en cuenta este efecto a la hora de interpretar datos (Charlesworth 1998; Noor y Bennett 2009), ya que puede hacernos interpretar que la recombinación es mutagénica (Hellmann *et al.* 2003), cuando no lo es, o identificar erróneamente regiones del genoma candidatas de haber sufrido selección reciente mediante el estadístico *Fst* (Charlesworth *et al.* 1997; Charlesworth 1998; Keinan y Reich 2010; Cutter y Choi 2010). Consultese Charlesworth 2012a para más detalles sobre esta problemática.

## 1.2 DE LA TEORÍA A LA PRÁCTICA: RETOS ACTUALES DE LA GENÉTICA DE POBLACIONES

En sus inicios la genética de poblaciones era un campo dominado por los modelos teóricos con escasos datos con los que ponerlos a prueba. En la actualidad, en la era de los genomas (véase cuadro 1), las elegantes herramientas matemáticas que se aplicaban sobre datos electroforéticos o secuencias fragmentarias del genoma se están reemplazando por nuevas y poderosas herramientas computacionales y algorítmicas para extraer, de nuevo, conocimiento a partir de genomas completos. La interacción entre procesos evolutivos (mutación, selección, deriva genética, migración, estructura poblacional y recombinación) genera tal complejidad en los patrones y dinámicas de la variación genética a lo largo de los cromosomas que hacen muy difícil su tratamiento analítico general (véase sección 1.3.3).

El objetivo de un genético de poblaciones actual es la estimación de las variables que ya sabemos determinan los patrones de variación genética a lo largo del genoma: (1) la tasa de recombinación, (2) la tasa de mutación, (3) el número de sitios bajo selección, (4) la *DFE*, (5) el grado de dominancia fenotípica de las mutaciones, (5) la historia demográfica de la población, (6) su estructura, (7) modo de apareamiento, y (8) ecología e historia natural para tantas especies como nos sea posible. No obstante, la inferencia de estas variables no es sencilla, pues distintas combinaciones de parámetros pueden dar lugar a los mismos patrones de variación genética (Ramírez-Soriano *et al.* 2008). Gracias a la disponibilidad de secuencias genómicas y a los avances en computación estamos empezando a ser capaces de estimar multitud de parámetros de modelos evolutivos muy complicados y poner en una balanza la importancia relativa de cada proceso evolutivo.

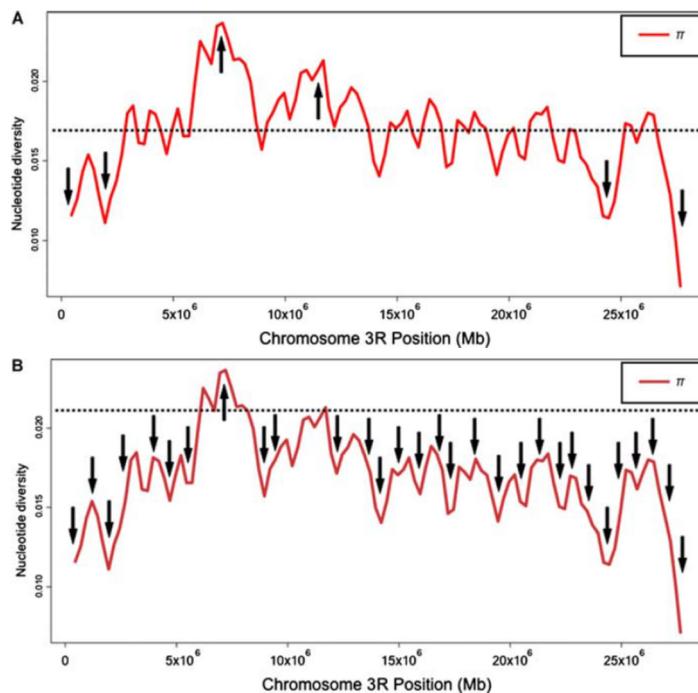
Como se comentó en la sección 1.1.3: Efecto Hill-Robertson, el hecho que alelos seleccionados no segreguen de manera independiente altera substancialmente las predicciones del modelo de Wright-Fisher para la probabilidad de fijación de

mutaciones seleccionadas (pero no de mutaciones neutras, véase cuadro 3). A su vez, las mutaciones seleccionadas alteran los niveles de variabilidad genética neutra ligada. De hecho, de aquí provienen las principales limitaciones de ambas teorías neutralistas. A continuación, se mostrará cómo la falta de segregación independiente entre mutaciones dificulta tanto la estimación de la huella de la selección natural como de la demografía cuando ambos procesos ocurren a la vez (sección 1.2.1). También se mostrará como la falta de segregación independiente entre sitios podría explicar, en parte, la escasa correlación entre el censo poblacional ( $N_c$ ) y los niveles de diversidad nucleotídica; esto es la denominada inicialmente por Lewontin (1974) paradoja de la variación (sección 1.2.2). Finalmente, se comentará qué factores afectan a la tasa y el tiempo de fijación de mutaciones adaptativas (sección 1.2.3).

### 1.2.1 DEMOGRAFÍA vs SELECCIÓN LIGADA

En paralelo al desarrollo de las Teorías Neutralistas (Kimura 1969a; Ohta y Kimura 1971), Cavalli-Sforza (1966) y Lewontin y Krakauer (1973) propusieron un criterio para diferenciar los efectos que tiene la selección y la deriva sobre la variación genómica el cual se ha convertido en un axioma de la genética de poblaciones moderna: “*While natural selection will operate differently for each locus and each allele at a locus, the effect of breeding structure is uniform over all loci and all alleles*” (Lewontin y Krakauer 1973). La idea viene a decir que mientras los efectos de la selección se circunscriben a una región pequeña del genoma, la historia demográfica de la población afecta al genoma entero. Muchos autores han utilizado este criterio para diferenciar los efectos de la selección y la demografía en grandes conjuntos de datos (Thornton *et al.* 2007). Bajo este enfoque se asume que la mayoría de genes aportan información sobre la historia demográfica de la población, mientras los genes que se encuentran en las colas de la distribución de algún estadístico (por ejemplo,  $n$  o *Tajima's D*) son las dianas más probables de la selección natural (figura 1.8A). Como todas las distribuciones tienen colas se han desarrollado mejoras en estos métodos que

proporcionen evidencias estadísticas de la acción de la selección natural. Esto puede hacerse simulando un amplio abanico de historias demográficas (Akey *et al.* 2004), simulando historias demográficas realistas provenientes de otras fuentes de datos (Stajich y Hahn 2005), o estimando la historia demográfica sobre el mismo conjunto de datos estudiado (Nielsen *et al.* 2005a).



**FIGURA 1.8** (A) La interpretación donde los sitios segregan independientemente, y (B) la interpretación donde la selección ligada afecta los niveles de polimorfismo. La línea roja en ambos casos muestra los valores de polimorfismo ( $\pi$ ) a lo largo del brazo cromosómico 3R de *D. simulans* (valores de Begun *et al.* 2007). Las flechas negras indican los efectos de la selección sobre los niveles de polimorfismo. La línea discontinua representa el valor esperado bajo el equilibrio mutación-deriva: en el panel A esto corresponde al valor promedio y en el panel B corresponde a un hipotético valor libre de los efectos de la selección ligada. [Figura tomada de Hahn (2008)].

La figura 1.8 trata de mostrar las diferencias entre la visión que asume que los sitios evolucionan de manera independiente (como hacen ambas teorías neutralistas), y la visión que contempla a la selección ligada, esto es la acción conjunta de barridos selectivos (*selective sweeps*) positivos y negativos como una fuerza a tener en cuenta para explicar los patrones de variación neutra. En la visión donde los sitios

evolucionan de manera independiente el criterio Cavalli-Sforza/Lewontin/Krakauer nos conduciría a pensar que el valor promedio de  $n$  representa el valor esperado bajo libre recombinación, independientemente de si la población está o no en el equilibrio demográfico, sólo los valores más extremos indicarían casos de selección equilibradora (valores elevados) o casos de selección positiva (valores bajos). La visión que considera que la selección ligada es omnipresente interpreta en cambio que la mayoría de los *loci* tienen niveles de  $n$  menores a los esperados bajo libre recombinación ya que los arrastres selectivos positivos y negativos disminuyen los niveles de variación y generan un exceso de alelos a baja frecuencia (Charlesworth *et al.* 1993; Stephan 2010).

De este modo, los *loci* con niveles altos de  $n$  podrían ser aquellos que han conseguido escapar completa o parcialmente de los efectos de la selección ligada gracias a una mayor tasa de recombinación local o a una menor densidad génica (aunque es posible que haya selección equilibradora genuina en algunos *loci* también). Debemos preguntarnos entonces si es posible estimar la demografía de las especies teniendo en cuenta que muy probablemente pocas regiones del genoma estarán libres de los efectos de la selección ligada (Hahn 2008; Li *et al.* 2012).

¿Qué evidencias tenemos a favor de la acción de la selección ligada que pongan en duda el criterio de Cavalli-Sforza/Lewontin/Krakauer? Bajo segregación independiente entre sitios no se espera ninguna correlación entre los niveles de polimorfismo y la tasa de recombinación (si esta es no mutagénica, véase más abajo) (Hudson 1983). Begun y Aquadro (1992) mostraron, sin embargo, que los genes localizados en regiones de baja recombinación mostraban niveles de polimorfismo más bajos que genes localizados en regiones de alta recombinación. Este resultado es hoy en día uno de los patrones más universales en genética de poblaciones, pues se ha encontrado una correlación positiva entre polimorfismo y recombinación en: humanos (Nachman *et al.* 1998; Przeworski *et al.* 2000), ratón (Nachman 1997;

Takahashi *et al.* 2004), *C. elegans* (Cutter y Payseur 2003), mosquito (Stump *et al.* 2005; Slotman *et al.* 2006), *A. thaliana* (Kim *et al.* 2007), tomate (Stephan y Langley 1998; Roselius *et al.* 2005), acelga (Kraft *et al.* 1998), maíz (Tenaillon *et al.* 2001), y trigo (Dvorak *et al.* 1998).

Existen dos hipótesis que podrían explicar esta correlación entre polimorfismo y recombinación: una neutra y otra selectiva. Si la recombinación es mutagénica, entonces regiones de alta recombinación tendrán más polimorfismo y más divergencia (Begun y Aquadro 1992). La hipótesis alternativa sugiere que la selección actúa a lo largo de todo el genoma y que las regiones con mayor recombinación es más probable que escapen de la selección (negativa o positiva) en sitios vecinos (Aquadro *et al.* 1994).

Analíticamente la reducción en la heterocigosidad neutra  $H_0$  esperada (bajo el modelo de Wright-Fisher en una población diploide de tamaño  $N$  con independencia entre sitios,  $H_0 = 4N\mu_0$ ) debido a la selección ligada ha sido estudiada y puede resumirse en la siguiente ecuación (Messer y Petrov 2013)

$$H_0 \approx 4N\mu_0 \times \frac{e^{-2\mu_d/r}}{1+8K(N)\nu s_b/r} \quad (1.6)$$

donde  $e^{-2\mu_d/r}$  es la reducción causada por la selección de fondo en la heterocigosidad neutra  $H_0$  esperada,  $\mu_d$  es la tasa de mutación deletérea por sitio ( $N_s < -1$ ) y  $r$  es la tasa de recombinación entre el sitio neutro y el sitio seleccionado (Hudson y Kaplan 1995; Stephan *et al.* 1999). Del mismo modo,  $(1 + 8K(N)\nu s_b/r)^{-1}$  es la reducción causada por arrastres positivos recurrentes en la heterocigosidad neutra  $H_0$  esperada, donde  $s_b$  es el coeficiente de selección de la nueva mutación beneficiosa,  $\nu$  es la tasa de mutación adaptativa por sitio y  $K(N)$  es una constante (Macpherson *et al.* 2007; Wiehe y Stephan 1993). Según esta ecuación hay una relación positiva entre recombinación y el nivel de variación neutra

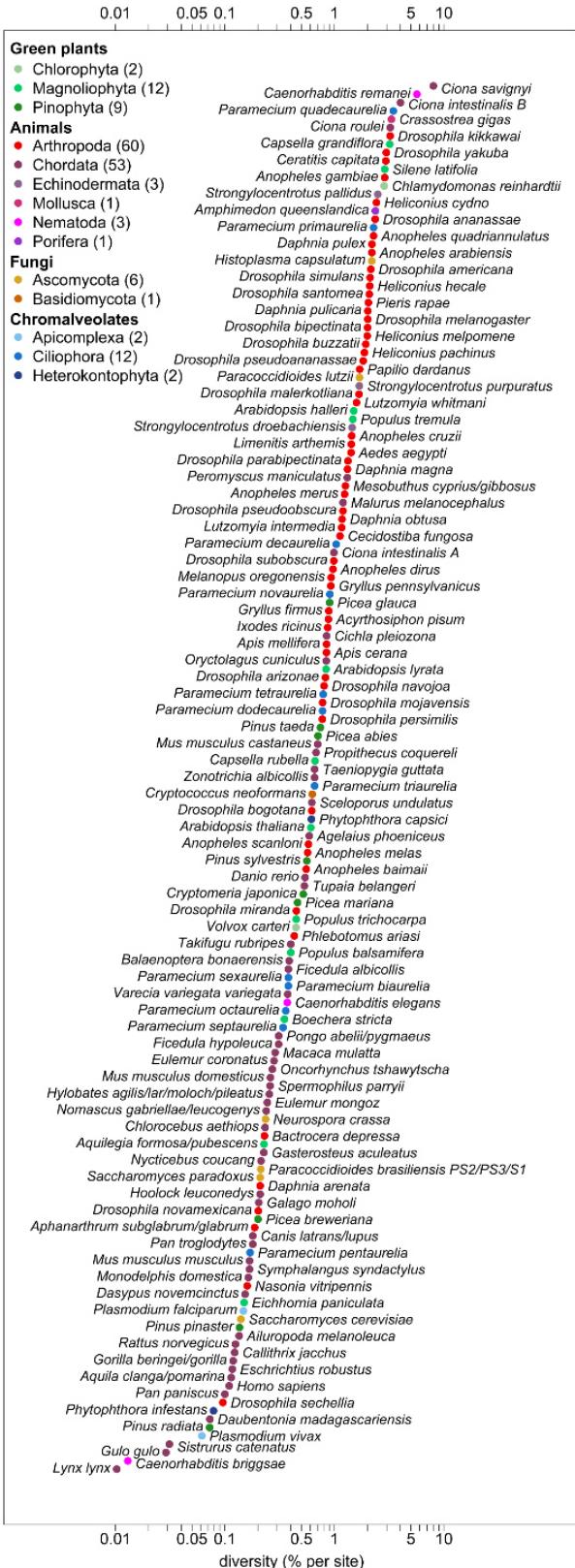
tal y como observamos en las poblaciones naturales de multitud de especies. Es importante tener en cuenta que en ausencia de otras fuerzas evolutivas la reducción en la variación causada por la selección ligada retorna a los niveles esperados bajo el equilibrio mutación-deriva relativamente rápido (Simonsen *et al.* 1995). El hecho que observemos una correlación entre polimorfismo y recombinación implica que en muchas especies han ocurrido arrastres selectivos (adaptativos o deletéreos) lo suficientemente recientes, intensos y/o abundantes como para que los niveles globales de polimorfismo no se encuentren en equilibrio mutación-deriva (Hahn 2008).

Es motivo de intenso debate que modo de selección ligada (selección de fondo o autoestopista – *hitchhiking*) es más importante para explicar los patrones de variación, ya que, ambos modelos predicen una correlación positiva entre polimorfismo y recombinación (Charlesworth 2012a; Barton 2010; Stephan 2010; Cutter y Payseur 2013). No obstante, muchos autores proponen incorporar la selección de fondo a las teorías neutralistas para tener un modelo nulo 2.0 que consiga mejorar nuestra interpretación de los patrones de variación genética a lo largo de los cromosomas y la detección de las regiones o genes del genoma que están detrás de la adaptación de las especies (Reed *et al.* 2005; McVicker *et al.* 2009; Hernandez *et al.* 2011; Lohmueller *et al.* 2011; Chun y Fay 2011; Charlesworth 2012a; Comeron 2014; Elyashiv *et al.* 2014; Corbett-Detig *et al.* 2015). En particular, el trabajo de Comeron (2014) muestra como no hay región en el genoma de *D. melanogaster* que esté libre de la selección de fondo. Este resultado tiene implicaciones negativas, como se ha comentado anteriormente, en nuestra capacidad de dar estimas realistas de la demografía, pero a su vez abre la puerta a poder detectar dianas recientes de selección equilibradora y/o selección positiva que antes, bajo el criterio de Cavalli-Sforza/Lewontin/Krakauer, no podían ser detectadas.

## 1.2.2 PARADOJA DE LA VARIACIÓN GENÉTICA

Comprender por qué algunas especies tienen más diversidad genética que otras es fundamental para el estudio de la ecología y la evolución, y tiene además importantes implicaciones para la biología de la conservación (Lynch y Lande 1998). Sin embargo, esta cuestión sigue sin resolverse. Con la rápida disminución de los costes de secuenciación (véase cuadro 1 y sección 1.1) se ha avanzado mucho en los últimos 5 años en lo que respecta la resolución de la que se ha denominado la paradoja de la variación genética (Lewontin 1974) que se planteó por primera vez hace 40 años gracias a datos de variación alozímica. Existe un intervalo misteriosamente estrecho de los niveles de diversidad genética neutra en diversas especies que varían notablemente en sus censos de población ( $N_c$ ).

De acuerdo con ambas teorías neutralistas de la evolución molecular (Kimura 1969a; Ohta y Kimura 1971), los niveles de diversidad genética neutra reflejan un equilibrio entre la entrada mediante nuevas mutaciones y la pérdida mediante la deriva de la diversidad genética (Kimura y Crow 1964; King y Jukes 1969; Kimura 1983). Bajo supuestos simplificadores, la intensidad de la deriva genética es inversamente proporcional al censo de población ( $N_c$ ). Los valores de diversidad en el equilibrio vienen dados por el producto de  $N_c$  y  $\mu$ , donde  $\mu$  es la tasa de mutación por generación. No obstante, las poblaciones fluctúan en tamaño a lo largo del tiempo y el éxito reproductivo de los individuos puede variar mucho (véase expresión [1.4]). A menudo, estas y otras desviaciones se pueden acomodar sustituyendo el censo de población  $N_c$  por un "tamaño efectivo de la población" o  $N_e$ , mucho más pequeño (Charlesworth 2009) (véase sección 1.1.1).



**FIGURA 1.9** Niveles de diversidad nucleotídica en cromosomas autosómicos en diferentes especies. El número promedio de diferencias entre todos los pares de secuencias por par de bases,  $n$  (Tajima 1983), se representa en % y en escala logarítmica en base 10. Cada valor representa la media de al menos tres *loci* para sitios sinónimos y/o ADN no-codificador. Las estimas están clasificadas por nivel de diversidad, etiquetadas de acuerdo a cada especie y coloreado por el filo al que pertenece cada especie. El número de especies en cada filo se indica entre paréntesis en la leyenda. [Figura tomada de Leffler *et al.* (2012)].

Si asumimos que la razón  $N_e/N_c$  es independiente de  $N_c$ , entonces esperamos observar una relación lineal con pendiente igual a 1 entre nuestras estimas de  $N_c$  (que pueden estar sujetas a un error importante) y nuestro estimador de  $N_e$  basado en la diversidad genética;  $n$  (Tajima 1983),  $n = 4N_e\mu$ , el cual depende también de la tasa de mutación por generación ( $\mu$ ). La figura 1.9 muestra los valores de  $n$  promedio en los “sitios” sinónimos (o no-codificadores), los cuales se asume que evolucionan de forma neutra, para genes autosómicos de 167 especies diferentes (Leffler *et al.* 2012). La paradoja surge, como se ha comentado antes, debido al estrecho intervalo de valores que encontramos de  $n$  respecto a los valores de  $N_c$ ; por ejemplo, la diferencia entre la

especie con mayor  $n$  respecto a la especie con menor  $n$  es de unas 800 veces (2-3 órdenes de magnitud), mientras la diferencia en  $N_c$  estará seguramente alrededor de los 8-10 órdenes de magnitud (según Leffler *et al.* 2012). En otras palabras, parece que una fuerza misteriosa hace que la razón  $N_e/N_c$  no sea independiente de  $N_c$ , de hecho, a mayor  $N_c$  menor  $N_e/N_c$  (y/o a menor  $N_c$  mayor  $N_e/N_c$ ).

Posibles determinantes de los niveles de variación neutra, y de  $N_e$ , no faltan. Los podemos agrupar en dos grandes categorías: (1) determinantes genético-poblacionales y (2) determinantes ecológicos y relacionados con la historia natural y biología de las especies. Evidencias recientes (Romiguier *et al.* 2014) parecen indicar que los determinantes relacionados con la historia natural de las especies aportan una pista importante para la resolución de la paradoja de la variación.

## DETERMINANTES GENÉTICO-POBLACIONALES

Tenemos la selección en sitios ligados como abanderados de este proceso; el *genetic draft* de Gillespie (2000a; 2000b; 2001) y la selección de fondo de Charlesworth (1993). Bajo ciertos supuestos, eventos de adaptación generalizada pueden limitar la gama de diversidad neutra entre especies: cuando la adaptación está limitada por la entrada de nuevas mutaciones, las poblaciones más grandes experimentan una mayor afluencia de nuevas mutaciones beneficiosas y, por tanto, mayores efectos de la selección ligada sobre los sitios neutros. A esta pérdida de variabilidad genética debido a arrastres selectivos recurrentes y fuertes se le denomina *genetic draft* (Gillespie 2001). En otras palabras, bajo ciertos supuestos, habrá más *genetic draft* en especies que experimentan menos deriva genética, y viceversa. El efecto combinado de *drift* y *draft* acabará estrechando el intervalo de diversidad neutra entre especies más de lo esperado por sus diferencias en  $N_c$  (Gillespie 2001). Sin embargo, el *genetic draft* requiere de la entrada de un torrente continuo de (nuevas) mutaciones fuertemente seleccionadas, y cabe la posibilidad que la fuente principal de

mutaciones adaptativas en poblaciones grandes provenga de mutaciones preexistentes (véase sección 1.2.3), con lo cual los niveles de variación no se verían tan drásticamente reducidos (Przeworski *et al.* 2005; Hermisson y Pennings 2005). Los primeros estudios (basados en contrastar los niveles de diversidad en diferentes regiones genómicas) sugirieron un impacto mucho más débil de la selección ligada sobre los niveles de variación neutra en humanos, *Drosophila* y diversas especies de plantas (Andolfatto 2007; Macpherson *et al.* 2007; Hernandez *et al.* 2011; Gossmann *et al.* 2011), descartando el *genetic draft* como explicación más plausible detrás de la paradoja. Confirmando esta observación, trabajos más recientes que tratan de explicar los niveles de variación neutra considerando la variación en la tasa de recombinación y la densidad génica a lo largo del genoma señalan que la selección de fondo es un modelo más plausible que el *genetic draft* o los arrastres selectivos clásicos (Maynard Smith y Haigh 1974), e incluso que el modelo neutralista clásico, para explicar los niveles de variación neutra hallados en *Drosophila* (Comeron 2014; Elyashiv *et al.* 2014; Corbett-Detig *et al.* 2015) y en otras 39 especies eucariotas diploides con reproducción sexual (Corbett-Detig *et al.* 2015). Es importante destacar que estos trabajos no contemplan otros modos de selección positiva como los *softs sweeps* y *partial sweeps* (véase sección 1.3.1) los cuales podrían, quizás, explicar mejor los patrones de variación neutra. En todo caso, el modelo de selección de fondo original contempla sólo mutaciones efectivamente seleccionadas ( $N_e s < -1$ ) y un tamaño de población constante (Charlesworth 1993). No obstante, si la fracción de mutaciones efectivamente deletéreas aumenta con el  $N_e$ , es lógico esperar que la selección de fondo sea más intensa (y por lo tanto más severa la reducción en  $n$ ) cuanto más grande sea la población. Este parece ser el caso de acuerdo a los resultados de Corbett-Detig *et al.* (2015) donde estimaron que la variación neutra perdida debido a la selección ligada (principalmente la selección de fondo) correlaciona positivamente con dos indicadores (*proxies*) del censo de población.

Datos recientes pueden ayudar a cuantificar el efecto de la selección ligada sobre la paradoja de la variación. Si existieran regiones del genoma libres de selección ligada, o si fuésemos capaces de estimar el valor máximo de heterocigosidad nucleotídica de sitios putativamente neutros ( $n$ ) en ausencia de selección ligada, estos valores libres de selección permitirían cuantificar el papel de la selección ligada en la paradoja de la variación (Leffler *et al.* 2012). Es decir, si los valores de  $n$  libres de selección ligada correlacionan de manera lineal y con pendiente igual a 1 con  $N_c$  (tal y como esperamos) la selección ligada explicaría en su totalidad la paradoja de la variación. Sorprendentemente, aunque Corbett-Detig *et al.* (2015) han sido capaces de calcular la  $n$  máxima en ausencia de selección ligada en 40 especies de animales y plantas (mediante modelos matemáticos que estiman la disminución en la variación debida a la selección de fondo y los arrastres selectivos positivos), los autores no correlacionan en ningún momento la  $n$  máxima con  $N_c$ , esto es sorprendente dado que así se podría cuantificar la contribución de la selección ligada a la paradoja de la variación. Coop (2016) reanalizando los datos de Corbett-Detig *et al.* (2015) encuentra una correlación débil entre la  $n$  libre de selección ligada y  $N_c$ , hecho que sugiere que la selección ligada es una fuerza menor a la hora de explicar las escasas diferencias observadas en los niveles de variación entre especies. Sin embargo, la selección ligada sigue siendo uno de los principales determinantes de la variación genética dentro de un mismo genoma.

Otra posible variable genética o intrínseca al genoma de las especies que puede ayudar a explicar la paradoja de la variación sería la existencia de una correlación negativa entre la tasa de mutación (nuclear) por generación y el tamaño efectivo,  $N_e$ . Esto podría deberse al aumento de la eficiencia de la selección en especies con censos efectivos grandes para disminuir la tasa de mutación,  $\mu$  (Lynch 2010; 2011). Estimaciones directas de la tasa de mutación por generación confirman esta hipótesis, la especie con la mayor tasa de mutación por generación estimada muta unas 100 veces más (2 órdenes de magnitud) que la especie con la menor tasa de

mutación por generación estimada (Leffler *et al.* 2012). Sin embargo, el aumento en el censo de población no escala perfectamente con la disminución de  $\mu$ ; pues como se ha dicho antes las diferencias en los censos de población están unos cuantos órdenes de magnitud por encima de las diferencias en  $\mu$ . La pregunta es entonces por qué no se cancelan estas dos variables la una a la otra para explicar enteramente la paradoja. La respuesta puede estar en la cantidad de ADN bajo selección (Sung *et al.* 2012); a igualdad de censo de población, genomas con mayor cantidad de ADN funcional pueden tener una mayor recompensa al disminuir  $\mu$ . En definitiva, cuantitativamente hablando este otro determinante genético no parece ser capaz de resolver por sí sólo la paradoja de la variación. No obstante, disponer de más estimas sobre la tasa de mutación en multitud de especies puede ayudar a corregir los niveles de diversidad y así poder correlacionarlos con el censo poblacional más limpiamente.

La selección en el uso de codones podría explicar también el estrecho intervalo de valores de  $n$ . Si una fracción substancial de las mutaciones sinónimas están débilmente seleccionadas ( $-10 < N_e s < -1$ ), la eficiencia de la selección purificadora será mayor en las poblaciones más grandes, y los niveles de diversidad aumentarán muy lentamente al aumentar  $N_e$  (Charlesworth y Charlesworth 2010; Ohta 1973). Sin embargo, un trabajo reciente muestra como la eficiencia de la selección en el uso de codones podría ser independiente del censo efectivo si existe un sesgo mutacional que contrarreste la dirección de la selección (Charlesworth 2013b). El sesgo mutacional conduce a un uso de codones fuera del óptimo, pero a medida que el censo efectivo aumenta de tamaño el uso de codones se acerca al óptimo y la fuerza de la selección disminuye. El resultado, la no correlación entre  $N_e$  y el carácter débilmente seleccionado (en este caso el uso de codones). Por lo tanto, de nuevo, resolver la paradoja invocando la selección débil en los sitios que asumimos neutros no parece la solución definitiva.

Para finalizar con los determinantes genéticos, destacar que constreñimientos estructurales en el apareamiento de los cromosomas podrían provocar que niveles muy altos de variación genética impidiesen el correcto apareamiento de los cromosomas y con ellos la segregación (Stephan y Langley 1992). Este constreñimiento impondría un límite superior de diversidad genética sólo para especies no haploides y podría contribuir también a la explicación del estrecho intervalo de valores de diversidad genética entre especies. A pesar de ello, no he encontrado ningún trabajo reciente que trate este tema de una manera sistemática, es decir, estudiando si realmente especies con censos poblacionales mayores son más susceptibles a tener problemas en la meiosis debido a una excesiva heterozigosidad.

## DETERMINANTES ECOLÓGICO-AMBIENTALES Y RELACIONADOS CON LA HISTORIA NATURAL Y BILOGÍA DE LAS ESPECIES

Las perturbaciones ambientales están detrás de las dinámicas y la diversidad de muchos ecosistemas, sin embargo, el papel de dichas perturbaciones en los patrones y distribución de la diversidad genética de las especies que habitan estos ecosistemas ha pasado desapercibido por la genética de poblaciones (Banks *et al.* 2013). No obstante, las perturbaciones ecológicas podrían imponer un techo superior a los valores de diversidad genética si estas fuesen más comunes en especies con censos de población grandes, lo que conduciría a dichas especies a experimentar de manera recurrente cuellos de botella más o menos extremos que alejarían sus niveles de diversidad genética de lo esperado bajo el equilibrio mutación-deriva (Kimura 1983; Haigh y Smith 1972; Leffler 2012). La pregunta que debemos hacernos si este resulta ser el caso es ¿qué conduce a dicha inestabilidad poblacional en las especies con grandes poblaciones?

Dentro de los caracteres ligados a la historia natural de las especies resulta obvio (para un genético de poblaciones) que el sistema de apareamiento de las especies tiene un gran impacto sobre la razón  $N_e/N_c$  (esto es, resumiendo, la razón entre los individuos que contribuyen a la descendencia en la siguiente generación respecto al total de individuos que componen la población) y los niveles de diversidad genética (Charlesworth y Wright 2001). Por ejemplo, la autofertilización reduce a la mitad el  $N_e$  esperado bajo panmixia (Caballero 1994; Nordborg 2000). Romiguier *et al.* (2014) encontraron que otros caracteres ligados a la historia natural o biología de las especies (menos relacionados a primera vista con la diversidad genética), como el tamaño de las crías (en cm) y su número, eran capaces de explicar alrededor del 60-70% de la varianza en la  $n$  sinónima para 76 especies de animales (pertenecientes a 29 familias). Esto supone un poder predictivo de la diversidad genética sin precedentes. Para simplificar, los autores clasifican a las especies que generan muchas crías de pequeño tamaño como *r* y a las especies que generan pocas crías de gran tamaño como *K*. Es decir, por un lado, tenemos a las especies longevas, de gran tamaño y con pocas crías (*K*) y por otro lado tenemos a las especies con esperanzas de vida cortas, pequeñas y que generan muchas crías (*r*). En medio encontraríamos una gama continua con características intermedias. ¿Qué hace que estas dos formas de enfrentarse a la vida sean tan importantes para determinar los niveles de variación genética? Los autores sugieren que estas estrategias podrían influir en la respuesta de las especies a las perturbaciones ambientales. Las especies tipo *K* han sido seleccionadas para la supervivencia y la optimización de la calidad de la descendencia en ambientes complejos y estables (MacArthur y Wilson 1968). Especulan que estas especies podrían experimentar menos perturbaciones ocasionales (o ser menos sensibles a ellas), lo que garantizaría la viabilidad a largo plazo de incluso pequeñas poblaciones. En cambio, sólo especies con poblaciones grandes son capaces de sostener la estrategia *r* ("riskier") a largo plazo, amortiguando así los cuellos de botella frecuentes experimentados en el contexto de alta sensibilidad ambiental. De acuerdo con esta hipótesis, las perturbaciones

ambientales serían un factor común que afectaría a todas las especies por igual, pero su impacto demográfico dependería de la estrategia de cada especie. En otras palabras, la paradoja de la variación podría explicarse en gran parte por el efecto combinado de las perturbaciones ambientales o ecológicas y la historia natural de las especies.

Esta colección de estudios demuestra que para resolver la paradoja de la variación no podemos descontextualizar a la genética de poblaciones ni de la ecología ni de la historia natural de las especies, a veces no es posible explicar los patrones de variación genética asumiendo que la ecología y la historia natural son variables aleatorias no relacionadas con parámetros poblacionales importantes (en este caso  $N_c$ ). En relación a estos estudios, es necesaria una nueva teoría que permita relacionar la ecología y la historia natural de las especies con los modos de selección y los patrones de variación genética a lo largo de distintas especies. Tales estudios no podrán proporcionar una respuesta universal, pero sí ayudarán a mejorar nuestra comprensión de la variación genética y sus determinantes.

### 1.2.3 TASA Y TIEMPO DE FIJACIÓN DE MUTACIONES ADAPTATIVAS

La adaptación es una cuestión fundamental en evolución molecular. Caracterizar el origen de las mutaciones beneficiosas es indispensable para comprender cómo las especies se adaptan a su entorno. Este tema sigue siendo un tremendo reto actualmente. Se han encontrado evidencias de selección positiva (en la sección 1.3 se explica en qué consisten los métodos estadísticos para detectar la selección positiva) tanto en el ADN no-codificador de *Drosophila* y ratones (Andolfatto 2005; Torgerson *et al.* 2009; Eyre-Walker y Keightley 2009; Kousathanas *et al.* 2011; Halligan *et al.* 2011; Halligan *et al.* 2013) como en el ADN codificador de *Drosophila*, ratones, bacterias y algunas plantas (Bustamante *et al.* 2002; Smith y Eyre-Walker 2002; Bierne y Eyre-Walker 2004; Sawyer *et al.* 2003; Charlesworth y Eyre-Walker

2006; Haddrill *et al.* 2010; Ingvarsson 2010; Slotte *et al.* 2010; Strasburg *et al.* 2009; 2011), a su vez hay escasas evidencias de selección positiva en genes codificadores de homínidos y algunas plantas (Chimpanzee Sequencing and Analysis Consortium 2005; Zhang y Li 2005; Boyko *et al.* 2008; Eyre-Walker y Keightley 2009; Gossman *et al.* 2010). En un trabajo reciente Galtier (2016) estima la tasa de adaptación proteica en 44 pares de especies de animales no-modelo, encuentra que la distribución de  $\alpha$  (esto es la proporción de substituciones adaptativas) es bimodal con un pico cercano a 0,3 y otro a 0,7. Prácticamente en todas las especies estudiadas encuentra evidencias de adaptación, con la interesante excepción de los homínidos. Las razones detrás de estas diferencias entre especies no se comprenden enteramente, aunque las diferencias en el  $N_e$  (Gossman *et al.* 2012; Galtier 2016) y la tasa de cambio de las condiciones ambientales entre especies (Razeto-Barry 2012; Lourenço 2013) podrían ser una explicación plausible. Además, el origen de la variación responsable de la adaptación (nótese que la adaptación puede ocurrir sobre variación preexistente y/o sobre nuevas mutaciones) también podría contribuir a explicar las diferencias observadas en la tasa de adaptación entre especies y sobretodo explicar ejemplos sorprendentes de adaptación rápida (Barret y Schlüter 2008).

### ***N<sub>e</sub>* VS TASA DE ADAPTACIÓN**

La teoría evolutiva predice que especies con grandes  $N_e$  deberían mostrar mayores tasas de evolución adaptativa que especies con  $N_e$  menores si asumimos que la adaptación está limitada por la entrada de nuevas mutaciones en la población (Kimura 1983). Esto es así porque en poblaciones grandes una mayor fracción de las nuevas mutaciones serán efectivamente beneficiosas y la tasa de substitución adaptativa es proporcional al producto  $\mu_a N_e s_a$  (donde  $\mu_a$  es la tasa de nuevas mutaciones adaptativas y  $N_e s_a$  es la fuerza de la selección positiva). La pendiente de la relación entre el  $N_e$  y la tasa de substituciones adaptativas dependerá del coeficiente de selección de las nuevas mutaciones beneficiosas  $s_a$  y la tasa de nuevas mutaciones adaptativas  $\mu_a$  (Ohta y Kimura 1971).

La adaptación está limitada por la tasa de mutación cuando la entrada de nuevas mutaciones beneficiosas es baja y hay períodos donde la población no contiene mutaciones beneficiosas (Lanfear *et al.* 2014). Especies con grandes censos efectivos (y/o tasas de mutación) podrían no sufrir esta limitación, porque estas tienden a producir muchas mutaciones (Cutter *et al.* 2013; Karasov *et al.* 2010) (consultese cuadro 4). Si la tasa de adaptación no está limitada por la entrada de nuevas mutaciones, entonces la relación entre  $N_e$  y la tasa de substituciones adaptativas dependerá de la relación entre los niveles de diversidad genética y el  $N_e$ . Sorprendentemente, como se ha comentado en la sección 1.2.2, los niveles de variación genética se encuentran pobremente correlacionados con el  $N_e$  (Lewontin 1974; Bazin *et al.* 2006; Nabholz *et al.* 2008; Leffler *et al.* 2012), lo cual sugiere que la relación entre el  $N_e$  y la tasa de adaptación podría ser muy pobre también (Bachtrog 2008; Karasov *et al.* 2010; Galtier 2016).

Otro factor que podría afectar la tasa de substituciones adaptativas entre especies es el efecto Hill-Robertson (Hill y Robertson 1966). En la práctica poco se sabe de cómo el  $N_e$  escala con la tasa de recombinación y la intensidad del efecto Hill-Robertson (Hill y Robertson 1966) entre especies. Sin embargo, algunos modelos teóricos sugieren que la frecuencia de los arrastres selectivos positivos (los cuales reducen la probabilidad de fijación de nuevas mutaciones beneficiosas y aumentan la de nuevas mutaciones deletéreas) aumenta con el  $N_e$  hasta que el número de arrastres alcanza un *plateau* causado por la interferencia clonal (Gillespie 2001). Otros modelos sugieren en cambio que la tasa de adaptación y los arrastres selectivos son independientes del  $N_e$  y que estos sólo dependen de la tasa de cambio de las condiciones ambientales y del número de caracteres bajo selección (Lourenço *et al.* 2013). En cualquier caso, ambos modelos predicen que la relación entre el  $N_e$  y la tasa de substituciones adaptativas debería ser prácticamente plana para valores elevados de  $N_e$  (Lanfear *et al.* 2014).

Hasta la fecha dos trabajos empíricos han evaluado la relación entre el  $N_e$  y la tasa de adaptación para un número considerable de especies: Gossman *et al.* (2012) y Galtier (2016). El primero ha encontrado una relación positiva débil, pero significativa, entre el  $N_e$  y la tasa de adaptación (estimada mediante el estadístico  $\omega_A$ , el cual es la razón entre la tasa de substituciones adaptativas y la tasa de substituciones neutras, véase sección 2.7) para 13 pares independientes de especies eucariotas. El segundo trabajo no encuentra, sin embargo, relación alguna entre el  $N_e$  y  $\omega_A$  para 44 pares independientes de especies animales. Trabajos basados en simulaciones respaldan los resultados de Galtier (2016) pues predicen también una falta de correlación entre el  $N_e$  y  $\omega_A$  (Razeto-Barry 2012; Lourenço 2013). Galtier (2016) explica esta falta de correlación aduciendo a una mayor tasa de nuevas mutaciones adaptativas ( $\mu_a$ ) compensatorias en especies con  $N_e$  menores. La tasa de nuevas mutaciones adaptativas sería mayor en poblaciones pequeñas porque estas deberían compensar la mayor fijación de mutaciones ligeramente deletéreas. En cambio, poblaciones grandes no fijan tantas substituciones ligeramente deletéreas y no se verían en la necesidad de compensar el efecto de dichas substituciones sobre la *fitness*. En otras palabras, aunque la fuerza de la selección ( $N_e s_a$ ) es menor en poblaciones pequeñas, la tasa de mutación adaptativa ( $\mu_a$ ) podría ser mayor en estas porque se encuentran más lejos del óptimo (Galtier 2016). La fuerza de la selección ( $N_e s_a$ ) y la tasa de mutación adaptativa ( $\mu_a$ ) se cancelan la una a la otra explicando la falta de correlación observada entre  $N_e$  y la tasa de adaptación entre especies (pues la tasa de adaptación es el producto  $\mu_a N_e s_a$ ).

## ADAPTACIÓN SOBRE VARIACIÓN PREEXISTENTE vs NUEVAS MUTACIONES

La velocidad de adaptación al medio (entendida como el tiempo que transcurre entre la llegada del cambio ambiental y la adaptación a dicho cambio por parte de todos los individuos de la población) es otra variable indispensable para entender cómo opera la evolución en poblaciones naturales. Trabajos teóricos han demostrado que

cuento la adaptación ocurre sobre variantes preexistentes (estas son variantes que están en más de una copia en la población) se produce una respuesta adaptativa más rápida al cambio ambiental, es decir el alelo adaptativo se fija antes, que cuando la adaptación ocurre sobre nuevas mutaciones, presentes en copia única (Hermisson y Pennings 2005). En otras palabras, son necesarias menos generaciones para que todos los individuos muestren la mutación beneficiosa. Además, no sólo la velocidad de adaptación aumenta, sino que, como es de esperar, el número de substituciones adaptativas que ocurren en un intervalo de tiempo dado también aumentan cuando la selección se da sobre variantes preexistentes (en comparación con lo esperado para nuevas mutaciones). Esto es así porque aumenta (sobretodo) tanto la probabilidad de fijación de alelos débilmente seleccionados (Hermisson y Pennings 2005) como de alelos parcialmente recesivos (Orr y Betancourt 2001; Hermisson y Pennings 2005), los cuales lo tienen muy difícil para fijarse cuando se encuentran en copia única. Consultese la revisión de Barret y Schlüter (2008) para algunos ejemplos paradigmáticos de adaptación rápida sobre variación preexistente en poblaciones naturales. Los arrastres positivos que ocurren sobre variación preexistente (o sobre mutaciones beneficiosas recurrentes, véase cuadro 4) producen una huella sobre la variación genética más débil que la producida por los arrastres clásicos o fuertes (*hard sweeps*) provenientes de nuevas mutaciones (Maynard Smith y Haigh 1974) – la variación se reduce menos cuando la adaptación se da sobre variantes preexistentes. A este tipo de arrastre alternativo se le denomina arrastre blando (*soft sweep*) (Hermisson y Pennings 2005) (véanse sección 1.3.1 y tabla 1.2 para más detalles).

Orr y Betancourt (2001) propusieron una forma de averiguar si la adaptación ocurre mayoritariamente sobre nuevas mutaciones o sobre variación preexistente. Esta prueba se basa en el contraste de la tasa de substitución adaptativa entre el cromosoma X y autosomas. Brevemente, como los machos son hemicigóticos para el cromosoma X, todas las mutaciones recesivas serán visibles para la selección en estos.

Esto produce un aumento de la tasa de evolución adaptativa en el X respecto a los autosomas siempre y cuando una fracción importante de las nuevas mutaciones beneficiosas sean completa o parcialmente recesivas. Estas se fijarán más fácilmente en el cromosoma X respecto a los autosomas (a este fenómeno se le conoce como evolución rápida del X [Charlesworth *et al.* 1987], en la sección 4.2.5 se explica en detalle). No obstante, la evolución rápida del X sólo se espera si la adaptación ocurre sobre nuevas mutaciones, si se da sobre variación deletérea preexistente (en equilibrio mutación-selección) se espera, en cambio, una mayor tasa de evolución adaptativa en autosomas porque las mutaciones deletéreas recesivas estarán a mayor frecuencia en los autosomas que en el X (véase Orr y Betancourt 2001 para conocer los detalles).

### VARIACIÓN DE LA TASA DE ADAPTACIÓN A LO LARGO DEL GENOMA

La tasa de evolución adaptativa también varía entre genes o regiones dentro de un genoma. Esto es así por razones diversas. Primero, algunos genes se esperan que desarrollen mayores tasas de adaptación por sus funciones biológicas, en particular los genes que interaccionan directamente con el mundo exterior, como los genes del sistema inmune implicados en el reconocimiento de patógenos y que se ven envueltos en carreras armamentísticas en contra de estos se espera que muestren mayores tasas de evolución adaptativa. En cambio, aquellos genes indispensables para el metabolismo o funciones de mantenimiento de la integridad y organización del genoma, como las histonas, se espera que se adapten muy lentamente o no muestren señales de adaptación reciente. Diversos estudios han mostrado que la función de los genes afecta a la tasa de evolución adaptativa: Obbard *et al.* (2009) han mostrado que los genes del sistema inmune tienen mayores tasas de adaptación que el resto de genes del genoma de *Drosophila*. A su vez, Haerty *et al.* (2007) y Pröschel *et al.* (2006) han mostrado que los genes con expresión sesgada en machos, como los genes presentes en los testículos o implicados en la generación del

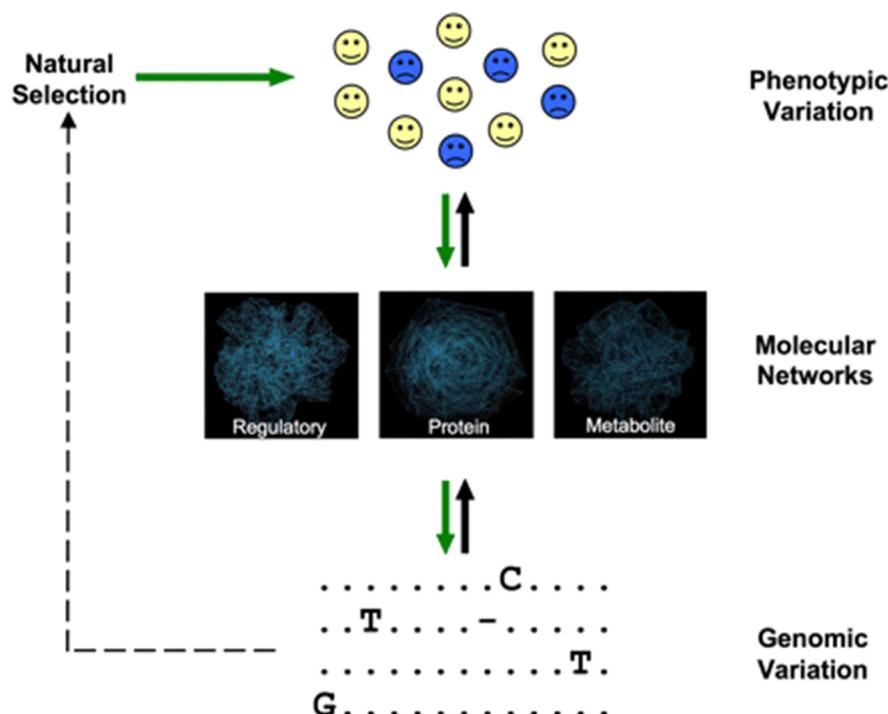
esperma, tienen elevadas tasas de adaptación en *Drosophila*. En humanos muchos genes que muestran señales de selección positiva tienden a estar sobrerepresentados en funciones como la percepción sensorial, sistema inmune, supresión de tumores, apoptosis y espermatozogénesis (Clark *et al.* 2003; Chimpanzee Sequencing and Analysis Consortium 2005; Nielsen *et al.* 2005b). Segundo, se espera que la tasa de evolución adaptativa dependa de la tasa de recombinación, genes en regiones de baja recombinación sufrirán la interferencia de Hill-Robertson (Hill y Robertson 1966; Felsenstein 1974) donde mutaciones seleccionadas interfieren unas con otras (véase sección 1.1.3). El papel de la recombinación sobre la tasa de adaptación ha sido previamente estudiado en *Drosophila* donde se ha encontrado una correlación positiva entre recombinación y adaptación (Presgraves 2005; Betancourt *et al.* 2009; Arguello *et al.* 2010; Mackay *et al.* 2012; Campos *et al.* 2014). Sorprendentemente, aunque también se espera que la variación en la densidad génica y la tasa de mutación a lo largo del genoma conlleve a una mayor o menor interferencia de Hill-Robertson, estas variables nunca se han correlacionado con la tasa de adaptación. Este es precisamente uno de nuestros objetivos en este trabajo.

### 1.3 ESTIMACIÓN DE LA HUELLA DE LA SELECCIÓN A LO LARGO DEL GENOMA

La genética de poblaciones ha evolucionado de ser un campo dominado por la teoría con escasos datos, hacia una disciplina dominada por los genomas donde conjuntos de datos genómicos ponen a prueba los modelos y los métodos de análisis computacionales disponibles. En la actualidad, en *Drosophila*, humanos, y otras pocas especies modelo, el análisis de los patrones de polimorfismo y divergencia a lo largo del genoma es la norma y no la excepción. Gracias al abaratamiento de los costes de secuenciación estos análisis podrán realizarse cada vez sobre más genomas de especies no modelo y sobre más individuos pertenecientes a dichas especies. Un objetivo central de la genética de poblaciones es determinar la fuerza y tasa de la selección natural sobre las mutaciones; este objetivo está cada vez más cerca gracias

a la cantidad de datos y herramientas estadísticas y computacionales disponibles.

Bajo el principio de la selección natural los caracteres beneficiosos son aquellos que aumentan la probabilidad de sobrevivir y reproducirse de los individuos, estos caracteres tenderán a ser más comunes en la población a lo largo del tiempo (Darwin y Wallace 1858; Darwin 1859). Haldane en 1949 fue el primero en proponer un fenotipo humano adaptativo, este era la resistencia endémica a la malaria en poblaciones con anemia (Haldane 2006). Años más tarde Allison (1954) encontró la base genética de dicha adaptación en el gen de la Hemoglobina-B.



**FIGURA 1.10** *Bottom-up population genomics*. Los test de selección a lo largo del genoma son agnósticos a los datos fenotípicos e infieren la selección en base a los patrones de variación genética (flecha negra discontinua). Sin embargo, la selección actúa directamente sobre la variación fenotípica y sólo indirectamente en la secuencia de ADN (flechas de color verde). Las flechas negras sólidas muestran que el camino desde la variación genética a la variación fenotípica transcurre a través de diversas redes moleculares dinámicas. [Figura tomada de Akey (2009)].

Este tipo de búsqueda de arriba a abajo (*top-bottom*), del potencial fenotipo adaptativo al genotipo responsable, se ha llevado a cabo con éxito en otras ocasiones: la persistencia de la lactasa y la pigmentación de la piel en humanos (Norton *et al.* 2007; Tishkoff *et al.* 2007), el color del pelaje en ratones (Mullen *et al.* 2009), y las espinas de los peces espinosos (Jones *et al.* 2012). Estos son ejemplos donde la genética de poblaciones sirve para probar hipótesis. Hoy en día los datos genómicos y los avances en las herramientas estadísticas permiten identificar las variantes genéticas candidatas a selección. Este hecho ha transformado a la genética de poblaciones en una ciencia generadora de hipótesis. A esta aproximación basada en la identificación de los genes candidatos a partir de los patrones de diversidad genética se le denomina comúnmente como de abajo a arriba (*bottom-up*) (figura 1.10).

Actualmente, proyectos de secuenciación de cientos (o miles) de individuos de una misma especie basados en las tecnologías de nueva generación (*next generation sequencing*) están permitiendo conocer cada vez más detalles de las variantes raras o a muy baja frecuencia, las cuales se espera tengan fuertes efectos sobre la varianza en la eficacia biológica entre individuos (Eyre-Walker 2010), y en el caso humano permitan entender mejor la arquitectura genética de muchas enfermedades (Eyre-Walker 2010). Es en humanos especialmente donde se están concentrando los mayores esfuerzos de secuenciación. Por ejemplo, el proyecto 1000 Genomes Project ha secuenciado a baja cobertura 2400 genomas completos en conjunto con la secuenciación de alta cobertura de los respectivos exomas (1000 Genomes Project Consortium 2010; 2012; 2015). También hay disponibles 44 genomas completos a alta cobertura (Shen *et al.* 2013) y el exoma completo para 6000 individuos (Fu *et al.* 2013). Doscientos dos genes han sido secuenciados a alta cobertura en 14000 individuos (Nelson *et al.* 2012). Es muy probable que datos de este tipo sigan acumulándose. El proyecto 100.000 Genomes Project pretende secuenciar dicha cantidad de genomas humanos completos en los próximos años (<http://www.genomicsengland.co.uk/the-100000-genomes-project/>). *D. melanogaster* es una de las especies modelo de la que

más datos genómico poblacionales disponemos, actualmente hay 623 genomas completos secuenciados a alta cobertura. El *Drosophila Genome Nexus* (Lack *et al.* 2015) es un recurso genómico poblacional con 623 genomas completos secuenciados a alta cobertura el cual ha reensamblado los 4 grandes conjuntos de datos genómico poblaciones disponibles (Langley *et al.* 2012; Mackay *et al.* 2012; Pool *et al.* 2012; Huang *et al.* 2014) y ha añadido más genomas de poblaciones sub-Saharianas. Este reensamblaje es necesario si pretendemos analizar el conjunto completo de genomas. En *Arabidopsis thaliana* hay más de 1000 individuos secuenciados (Nordborg *et al.* 2005; Gan *et al.* 2011; <http://1001genomes.org>) y el genoma completo de 40 gusanos de seda también puede utilizarse (Xia *et al.* 2009). Antes de disponer de estos datos provenientes de las tecnologías *next-generation*, otros estudios genómico poblacionales se realizaron mediante la secuenciación de Sanger (Bustamante *et al.* 2005; Begun *et al.*, 2007) o genotipado de SNPs (Hinds *et al.* 2005; International HapMap Consortium 2005, 2007; Jakobsson *et al.* 2008; Li *et al.* 2008). La secuenciación a baja cobertura de seis genomas de *Drosophila simulans* por Begun *et al.* (2007) supuso el primer estudio de la genómica de poblaciones propiamente dicho (Hahn 2008), sin embargo, hoy en día un solo *run* de *Illumina Genome Analyzer* puede producir substancialmente más datos de los que estaban presentes en ese estudio. Aparte de estos avances substanciales en lo que a datos genómicos poblacionales se refiere, también hemos acumulado datos relacionados con la regulación, nivel de expresión génica, localización de las regiones reguladoras y estados cromatínicos en *Drosophila*, *C. elegans* y humanos en distintos tejidos y etapas del desarrollo, datos puestos a disposición de la comunidad científica por los proyectos modENCODE (The modENCODE Project Consortium 2010) y ENCODE (The ENCODE Project Consortium 2012), respectivamente. La base de datos REDfly se encarga, además, de verificar experimentalmente los elementos reguladores del genoma de *Drosophila* (Gallo *et al.* 2011). Estos recursos permiten caracterizar mejor las dianas de la selección natural y la relación o mapa fenotipo-genotipo.

En conclusión, durante los últimos 15 años hemos vivido, en paralelo a los avances en secuenciación, el desarrollo y aplicación de numerosos métodos estadísticos para identificar las regiones genómicas que presentan la huella de la selección natural. Estos métodos han sido utilizados para investigar eventos selectivos de distintas edades en diversas especies. Las mejoras en la anotación de los genomas, de su manipulación, y su interrogación mediante técnicas de biología molecular masivas u ómicas, están permitiendo además cruzar la frontera de la identificación de las variantes candidatas, a su caracterización funcional y su papel en la evolución (véase figura 1.10). En esta sección se muestran brevemente muchos de estos métodos y se describe en más detalle los que han sido más importantes para este trabajo.

### 1.3.1 TIPOS DE SELECCIÓN NATURAL

Existen muchos modelos matemáticos que tratan de describir los distintos modos conocidos de selección natural y sus efectos y cuantificación a nivel molecular (revisiones recientes en Pool *et al.* 2010; Crisci *et al.* 2012; Cutter y Payseur 2013). Algunos tipos o modos de selección solapan conceptualmente y otros son directamente equivalentes. En la era genómica, la selección se define como la propagación diferencial de un alelo como consecuencia de su efecto sobre la eficacia biológica (Akey 2009).

El modo más sencillo de selección es aquel en el que la selección actúa en una dirección, si el alelo se ve favorecido y se propaga hablamos de selección positiva, si por el contrario el alelo está desfavorecido hablamos de selección negativa o purificadora. Cuando en un mismo *locus* dialélico ambos alelos se mantienen a una frecuencia apreciable debido a la acción de la selección natural, denominamos a este fenómeno selección equilibradora. Estos casos pueden ocurrir por ejemplo cuando el heterocigoto es superior a los dos homocigotos; si este es el caso podemos hablar también de ventaja del heterocigoto (*heterozygote advantage*) o sobredominancia

(*overdominance*). La selección equilibradora también puede ocurrir cuando el coeficiente de selección del alelo cambia dependiendo de su frecuencia, por ejemplo, el alelo puede ser adaptativo sólo cuando está a baja frecuencia (Charlesworth 2006). A este tipo de selección que depende de la frecuencia se le denomina en los libros de texto y artículos como selección dependiente de las frecuencias (*frequency-dependent selection*). Cuando el ambiente donde vive la población cambia rápidamente podemos encontrarnos casos donde en un mismo *locus* el alelo *A* es beneficioso en la situación *X*, pero deletéreo en la situación *Y*, y lo contrario para el alelo *a*. A este escenario se le denomina sobredominancia marginal (*marginal overdominance*) y recientemente ha sido motivo de estudio en *D. melanogaster* donde se han identificado cientos de SNPs que sufren cambios dramáticos en su frecuencia (entre 40-60%) estacionalmente año tras año (Bergland *et al.* 2014).

A pesar de esta diversidad de modos de selección, la mayoría de la investigación se ha centrado en el desarrollo de métodos para detectar la selección positiva, en concreto los barridos fuertes (*hard sweeps*, véase más abajo). Tanto razones prácticas como teóricas explican este sesgo. La señal o huella que deja tras de sí la selección positiva es más fácil de detectar que la señal producida por la selección purificadora o la selección equilibradora. Desde el punto de vista teórico, la selección positiva es el mecanismo primario de la adaptación, es decir, de la génesis de los fenotipos que aumentan la eficacia biológica de los individuos ante nuevos ambientes o nichos ecológicos (Akey 2009). Centrándonos en la selección positiva esta puede ocurrir sobre nuevas mutaciones, dando lugar a los denominados arrastres clásicos o barridos fuertes (Maynard Smith y Haigh 1974). A los barridos fuertes que están en proceso de fijación se les denomina como barridos parciales (*partial sweeps*, Voight *et al.* 2006). Los barridos parciales pueden producirse también si el alelo beneficioso deja de serlo. Assaf *et al.* (2015) han descrito un nuevo tipo de arrastre que denominan arrastres escalonados (*staggered sweep*) en el cual la variante seleccionada positivamente se encuentra ligada a otra variante fuertemente deletérea recesiva. En

en este escenario el haplotipo empieza a aumentar de frecuencia y después se mantiene a cierta frecuencia hasta que la recombinación separa ambos *loci* seleccionados. Cuando esto ocurre el alelo beneficioso acaba de fijarse y el deletéreo disminuye drásticamente de frecuencia. Es un arrastre escalonado porque se da en dos fases. Además, la selección positiva puede ocurrir sobre mutaciones preexistentes (*standing genetic variation*) que antes eran neutras o deletéreas, dando lugar a los barridos suaves (*soft sweeps*, Hermisson y Pennings 2005; Barret y Schlüter 2008) (véase sección 1.2.3). No obstante, los barridos suaves pueden darse también cuando la misma mutación beneficiosa ocurre de manera independiente sobre distintos cromosomas en un breve intervalo de tiempo (Karasov *et al.* 2010) (cuadro 4). Esta segunda versión de barido suave se ha contemplado tradicionalmente como algo muy raro, ya que para que se dé son necesarios censos efectivos y/o tasas de mutación tan grandes como para permitir que una misma mutación ocurra dos veces en una misma generación o en pocas generaciones de diferencia. En el cuadro 4 se considera un caso real de convergencia adaptativa en la que una misma mutación ocurre independientemente en *Drosophila*. Este ejemplo parece contradecir nuestra concepción de la relación entre la adaptación y  $N_e$ .

Todos los casos comentados de selección se centran en relaciones uno a uno entre el genotipo y el fenotipo, es decir, un gen un carácter, sin embargo, las relaciones donde múltiples genes determinan un mismo fenotipo, o donde un mismo gen puede afectar a distintos fenotipos se espera sean más comunes (Falconer y Mackay 1996) (en este contexto cuando hablo de genes en realidad me refiero a mutaciones). La adaptación poligénica se utiliza para describir el proceso en el que la adaptación se produce por la selección simultánea de alelos en muchos *loci* (Barton 1989; Pritchard *et al.* 2010; Pritchard y Di Rienzo 2010). La adaptación poligénica puede ocurrir cuando un fenotipo cuantitativo se desplaza ligeramente de su óptimo, en este escenario muchos alelos, o variación preexistente, de pequeño efecto podrá cambiar de frecuencia sin dejar apenas señal de arrastre. La adaptación poligénica puede ocurrir

también sobre nuevas mutaciones en muchos *loci* si el fenotipo recién favorecido había estado previamente fuertemente desfavorecido (Pritchard *et al.* 2010). Las señales producidas por este tipo de selección son difícilmente diferenciables del trasfondo neutro (Barton 1989; Pritchard y Di Rienzo 2010).

#### CUADRO 4: LA TASA DE MUTACIÓN NO LIMITA LA ADAPTACIÓN A LA RESISTENCIA EN PESTICIDAS EN *D. melanogaster*

El trabajo de Karasov *et al.* (2010) es un ejemplo paradigmático de cómo se aplica el razonamiento de genética de poblaciones en el análisis de datos nucleotídicos. Observaron que la misma mutación que daba resistencia a pesticidas había ocurrido independientemente en distintas poblaciones de *D. melanogaster* en un periodo de tiempo muy corto. Supieron que eran mutaciones independientes, a pesar que el cambio de nucleótido era exactamente el mismo, porque el haplotipo donde había ocurrido era distinto en cada población estudiada. Este hecho es, sin embargo, incompatible con los niveles de variación genética observados. Es decir, dado el producto de la tasa de mutación de *D. melanogaster* ( $\mu \sim 3,5 \times 10^{-9}$  [Keightley *et al.* 2009]) por su censo efectivo ( $N_e \sim 10^6$  [Thornton y Andolfatto 2006; Li y Stephan 2006]) la aparición de la misma mutación en un intervalo tan corto de tiempo es altamente improbable, ya que,  $\Theta = 4N_e\mu \ll 1$ , por no decir imposible (Karasov *et al.* 2010).

Sin embargo, este tamaño efectivo es sólo una descripción de los niveles de diversidad neutra, no nos dice cómo la deriva genética influye a la adaptación más reciente. Es crucial distinguir aquí entre los factores que disminuyen  $N_e$  a corto plazo (*short-term N<sub>e</sub>*), tales como la variación en el número de crías por progenitor o el efecto Hill-Robertson (véase sección 1.1.3), los cuales aumentan el efecto de la deriva a nivel genómico o a nivel de alelo, respectivamente, y acontecimientos más drásticos como cuellos de botella demográficos o arrastres selectivos recurrentes que reducen la diversidad neutra y son indicadores de un  $N_e$  histórico o a largo plazo (*long-term N<sub>e</sub>*). En otras palabras, factores que afectan a la eficiencia de la selección a nivel global (en todo el genoma) como la variación en el número de crías o la proporción de ambos sexos pueden reducir el  $N_e$  a corto plazo hasta un orden de magnitud por debajo del censo poblacional  $N_c$  (Frankham 1995), la interferencia de Hill-Robertson puede a su vez reducir localmente el  $N_e$  a corto plazo hasta dos órdenes de magnitud por debajo del  $N_e$  determinado por los factores que actúan a escala genómica (Messer y Petrov 2013, Charlesworth 2013a). En cambio, los arrastres selectivos y cuellos de botella intensos son *a priori* independientes del censo poblacional ( $N_c$ ) y reducen el censo efectivo ( $N_e$ ) a largo plazo (véase la paradoja de la variación en la sección 1.2.2 para más detalles).

Karasov *et al.* (2010) afirman que el  $N_e$  de las últimas 1000-1500 generaciones (o 50 años) es  $\gg 10^8$ , hecho que posibilitaría que en una misma generación se produjese la misma mutación más de una vez sobre distintos cromosomas, ya que,  $\Theta = 4N_e\mu \gg 1$ . ¿Cómo reconciliar este nuevo  $N_e$  que explicaría el patrón de adaptación observado con el  $N_e$  estimado a partir de los niveles de variación nucleotídica neutra? En poblaciones fluctuantes, el tamaño efectivo es la media armónica de los tamaños efectivos en generaciones individuales (Charlesworth 2009) y por lo tanto su valor actual está dominado por los valores más pequeños de  $N_e$ . Las estimas de  $N_e$  basadas en los niveles de variación nucleotídica neutra reflejan la media armónica del  $N_e$  durante largos períodos de tiempo y por lo tanto es muy sensible a los períodos de bajo  $N_e$  (Lewontin 1974). Además, Karasov *et al.* (2010) demuestran mediante simulaciones que añadir un tamaño poblacional actual de  $10^8$  a la demografía de *Drosophila* no afecta significativamente las estimas de variación neutra actuales.

#### **CUADRO 4: CONTINUACIÓN**

Este resultado tiene importantes implicaciones, ya que parece sugerir que la tasa de adaptación no se ve limitada por la tasa de mutación. Esto es cierto sólo parcialmente. Las mutaciones ligeramente beneficiosas ( $1 < N_e s < 10$ ) requieren más tiempo para fijarse y estas serán más susceptibles a sufrir los arrastres selectivos y cuellos de botella demográficos que las mutaciones fuertemente beneficiosas ( $N_e s > 10$ ), las cuales se fijarán en las etapas de *boom* demográfico (que podrían ser las más comunes). Es decir, es posible que el  $N_e$  a largo plazo estimado a través de la variación neutra actual sea un buen indicador del  $N_e$  que es relevante para las mutaciones ligeramente beneficiosas, pero no para las mutaciones fuertemente beneficiosas las cuales dependerán más de los períodos de *boom* demográfico. En definitiva, si la mayoría de las mutaciones beneficiosas lo son fuertemente, la tasa de adaptación será independiente de la tasa de mutación, si por el contrario la mayoría de mutaciones beneficiosas son ligeramente beneficiosas (como apuntan Schneider *et al.* 2011 en *Drosophila*), entonces la tasa de adaptación dependerá de la tasa de mutación.

### 1.3.2 PRUEBAS DE LA TEORÍA NEUTRALISTA (*NEUTRALITY-TESTS*)

Los genéticos de poblaciones han tratado durante mucho tiempo de cuantificar la importancia relativa de las distintas fuerzas evolutivas en los patrones de variación genética de las poblaciones naturales (Lewontin 1974). Este objetivo descansa sobre tres pilares: teoría, observación experimental e inferencia estadística que enlace la observación experimental con la teoría. La genética de poblaciones teórica describe matemáticamente como las distintas fuerzas evolutivas interaccionan para producir distintos patrones de variación genética dentro y entre especies. La genética de poblaciones experimental trata de cuantificar los niveles de variación genética en un grupo dado de poblaciones naturales e interpretar esta variación de acuerdo a la teoría. El problema de esta tarea es que los patrones observados de variación pueden ser igualmente explicados invocando distintas fuerzas evolutivas o interacciones entre estas. En términos estadísticos, hay demasiados grados de libertad. La solución estándar ha sido asumir un modelo nulo muy estricto, este es la Teoría Neutra original de Kimura (1983) (sección 1.1.2). Gran parte de los esfuerzos estadísticos en genética de poblaciones se han basado en el diseño de pruebas de neutralidad.

Las pruebas de la teoría neutralista (*neutrality-test*) consisten en estadísticos que comparan los niveles y patrones observados de variación con los esperados bajo el modelo nulo, el cual es normalmente la versión más ortodoxa de la teoría neutralista (Kimura 1983). A pesar de todas sus limitaciones la teoría neutralista permite desarrollar test de selección basados o bien en la simulación de los patrones de variación genética generados por la deriva (y la demografía), o bien en los patrones de diversidad genética observados en sitios que se asume evolucionan de forma neutra (generalmente las posiciones sinónimas). Existen aproximaciones heurísticas, correcciones o extensiones *ad hoc* de algunos de estos estimadores puntuales para casos donde la teoría neutralista no se cumple estrictamente. En la tabla 1.2 se listan los estimadores puntuales más importantes y sus extensiones heurísticas, o basadas en inferencia de parámetros mediante máxima verosimilitud, más populares.

**TABLA 1.2 MÉTODOS PARA LA DETECCIÓN DE LA SELECCIÓN NATURAL A NIVEL MOLECULAR**

Edad del evento selectivo	Datos	Características	Test Representativos	Referencias
Macroevolución (entre especies)	Divergencia	Generalmente se aplica a secuencias codificadoras. Las substituciones sinónimas se consideran neutras. Si la tasa de sustitución no-sinónima es mayor a la tasa de sustitución sinónima ( $\omega > 1$ ) invocamos a la selección positiva. Si $\omega = 1$ , los sitios no-sinónimos evolucionan de manera neutra y si $\omega < 1$ la selección purificadora domina. Test muy conservador a nivel de proteína pero mejora si se aplica a nivel de codón en arboles filogenéticos lo suficientemente informativos. Existen extensiones donde se estima el sesgo en el uso de codón, el sesgo mutacional transición/transversión y la intensidad de la selección por sitio.	$K_a/K_s$ ( $dN/dS$ o $\omega$ )	Graur y Li (2000); Hurst (2002); PAML (Nielsen y Yang 1998; Yang 1997)
		Puede aplicarse a cualquier región. Regiones conservadas en muchas especies pero que en un linaje muestra más substituciones de las esperadas son candidatas a selección positiva.	Identificación de regiones de evolución acelerada	Burbano <i>et al.</i> (2012); Lindblad-Toh <i>et al.</i> (2011); Pollard <i>et al.</i> (2006); Prabhakar <i>et al.</i> (2006); Shapiro y Alm (2008)
	Divergencia y Polimorfismo	Generalmente se aplica a secuencia codificadora (pero se puede extender a no-codificador). Compara la razón entre el número de substituciones no-sinónimas ( $d_N$ ) respecto el número de substituciones sinónimas ( $d_S$ ) y la razón entre el número de polimorfismos no-sinónimos ( $p_N$ ) respecto el número de polimorfismos sinónimos ( $p_S$ ). Si $d_N/d_S > p_N/p_S$ invocamos selección positiva. Si $d_N/d_S = p_N/p_S$ invocamos selección purificadora. Si $d_N/d_S < p_N/p_S$ invocamos selección equilibradora o a un exceso de alelos no-sinónimos ligeramente deletéreos a baja frecuencia. Existen tres extensiones heurísticas del MKT que le permiten corregir la presencia de alelos ligeramente deletéreos segregando en la población y/o de la falta de independencia entre sitios: (1) Corrección de Fay, (2) <i>Direction of Selection (DoS)</i> y (3) <i>Asymptotic MKT</i> . Véase texto principal.	Prueba de McDonald-Kreitman (MKT)	McDonald y Kreitman (1991); Egea <i>et al.</i> (2008); Fay <i>et al.</i> (2001; 2002); Stoletzki y Eyre-Walker (2011); Messer y Petrov (2013)
		Puede aplicarse a cualquier región. La razón entre la divergencia (D) y el polimorfismo (P) entre <i>loci</i> depende de la tasa de mutación, $\mu$ . Si D/P para el <i>locus</i> 1 es $> D/P$ para el resto de <i>loci</i> invocamos la selección positiva. Si D/P para el <i>locus</i> 1 es $< D/P$ para el resto de <i>loci</i> invocamos la selección balanceada. Hay extensiones para calcular el valor D/P para un linaje mientras se tiene en cuenta la variación en $\mu$ .	Prueba de Hudson-Kreitman-Aguadé (HKA)	Hudson <i>et al.</i> (1987); Wright y Charlesworth (2004)

Microevolución $(< 4N_e$ generaciones promedio en diploides)	Espectro de Frecuencias	En un arrastre selectivo fuerte ( <i>hard sweep</i> ), el alelo seleccionado alcanza alta frecuencia en la población junto con variantes ligadas a este. Esto genera un exceso de variantes derivadas a alta frecuencia pues la recombinación rompe el haplotípico a medida que este se fija. Las nuevas mutaciones que ocurren sobre este fondo homogéneo aumentan la fracción de alelos a baja frecuencia.	Ewens-Watterson Test	Ewens (1972); Watterson (1978)
			Tajima's D y derivados	Tajima (1989); Tajima (1993); Fu (1997); Fu y Li (1993)
			Fay y Wu's H	Fay y Wu (2000); Fay (2011)
			SweepFinder y derivados	Nielsen <i>et al.</i> (2005a); Pavlidis <i>et al.</i> (2013)
	Desequilibrio de Ligamiento	En un arrastre selectivo fuerte ( <i>hard sweep</i> ) (completo o parcial), el alelo seleccionado alcanza alta frecuencia en la población junto con variantes ligadas a este. Es decir, en la región donde ha ocurrido el arrastre observamos un único haplotípico muy largo y una diversidad haplotípica muy baja.	<i>Long-range haplotype test</i> (LRH)	Sabeti <i>et al.</i> (2002); Zhang <i>et al.</i> (2006)
			<i>Long-range haplotype similarity test</i>	Hanchard <i>et al.</i> (2006)
			<i>Integrated haplotype score</i> (iHS)	Voight <i>et al.</i> (2006)
			<i>Cross-population extended haplotype homozygosity</i> (XP-EHH)	Sabeti <i>et al.</i> (2007)
			<i>Linkage disequilibrium decay</i> (LDD)	Wang <i>et al.</i> (2006)
			<i>Identity-by-descent (IBD) analyses</i>	Cai <i>et al.</i> (2011); Han y Abney (2012)
	Diferenciación Poblacional	La selección que actúa sobre un alelo en una población pero no en otra genera una diferencia en la frecuencia del alelo entre las dos poblaciones mayor a la esperada para un alelo neutro.	<i>Haplotype homozygosity</i> (H12 y H2/H1)	Garud <i>et al.</i> (2015)
			Lewontin-Krakauer test (LKT)	Lewontin y Krakauer (1973); Vitalis <i>et al.</i> (2001); Excoffier <i>et al.</i> (2009); Bonhomme <i>et al.</i> (2010)
			<i>Locus-specific branch length</i> (LSBL)	Shriver <i>et al.</i> (2004)
			hapFLK	Fariello <i>et al.</i> (2013)

[Tabla adaptada y ampliada a partir de Vitti *et al.* (2013)].

## PRUEBA DE MCDONALD-KREITMAN EN PRESENCIA DE ALELOS LIGERAMENTE DELETÉREOS, CAMBIOS DEMOGRÁFICOS Y SELECCIÓN LIGADA

La prueba de McDonald y Kreitman, MKT (1991) es un poderoso test de la evolución neutra a nivel molecular, además puede utilizarse para estimar la fracción de substituciones fijadas a través de la selección positiva (Charlesworth 1994; Akashi 1999; Fay *et al.* 2001; Smith y Eyre-Walker 2002), mediante el estadístico  $\alpha$

$$\alpha = 1 - \frac{D_s}{D_n} \frac{P_n}{P_s} \quad (1.7)$$

En el MKT los niveles de variación dentro de una especie (polimorfismo,  $P$ ) son comparados con los niveles de variación entre especies (divergencia,  $D$ ) para dos tipos de sitios. Normalmente estos sitios son las posiciones cero veces degeneradas o no-sinónimas ( $n$ ) y las posiciones cuatro veces degeneradas o sinónimas ( $s$ ). Sin embargo, el test puede aplicarse a otras clases de sitios (Jenkins *et al.* 1995; Akashi 1995; Kohn *et al.* 2004; Andolfatto 2005; Casillas *et al.* 2007; Egea *et al.* 2008; Mackay *et al.* 2012).

Bajo la hipótesis nula (esto es la teoría neutralista original, véase figura 1.3B), se espera que todas las mutaciones sinónimas sean neutras y que las mutaciones no-sinónimas sean o bien neutras, o bien fuertemente deletéreas o fuertemente beneficiosas. En cualquier caso, se espera que todos los alelos no-sinónimos sean neutros, pues las mutaciones fuertemente seleccionadas permanecen poco tiempo segregando en la población. La razón  $P_n/P_s$  es una medida del constreñimiento selectivo en la fase polimórfica y se asume que este es equivalente al constreñimiento en la fase divergente, es decir, en ausencia de adaptación  $P_n/P_s = D_n/D_s$ . Si  $D_n/D_s > P_n/P_s$  entonces hay un exceso de  $D_n$  debido a la fijación de mutaciones beneficiosas. Aunque podría ser que el constreñimiento haya aumentado recientemente, o que el constreñimiento haya sido menor en el pasado, creando evidencias de selección positiva artefactuales y poniendo en duda la aplicabilidad del MKT (Nei *et al.* 2010; Fay 2011).

Tal y como apuntaron McDonald y Kreitman (1991) en su trabajo original, los cambios demográficos podrían explicar la fluctuación temporal del constreñimiento selectivo y con ello la sobreestima del papel de la selección positiva en la tasa de evolución. Este argumento fue estudiado en detalle por Eyre-Walker (2002) y requiere de la presencia de alelos ligeramente deletéreos ( $-10 < N_e s < -1$ ). Durante un período pasado de bajo  $N_e$ , mutaciones ligeramente deletéreas podrían haberse fijado. En la actualidad estas mutaciones no segregan debido a que la población tiene un mayor  $N_e$ , la selección las elimina más eficientemente. Por lo tanto, estas contribuyen a  $D_n$  pero no a  $P_n$ , y  $\alpha > 0$  incluso en ausencia de substituciones adaptativas. Una estrategia para minimizar este problema es elegir poblaciones que sepamos no hayan sufrido cambios demográficos severos y prolongados (algo muy complicado de conocer). Otra alternativa es utilizar los métodos disponibles que estiman la historia demográfica reciente junto con  $\alpha$  (Boyko *et al.* 2008; Eyre-Walker y Keightley 2009; Schneider *et al.* 2011). Véase sección 1.3.3 para más detalles. No obstante, estos métodos siguen sin poder controlar para eventos demográficos lejanos de los que ya no queda constancia en el polimorfismo.

Las extensiones o correcciones heurísticas para el MKT se han propuesto cuando hay desviaciones de los supuestos de la teoría neutralista, esto es, cuando mutaciones ligeramente deletéreas segregan en la población (Fay *et al.* 2001; Charlesworth y Eyre-Walker 2008) y el genoma está sometido a arrastres selectivos recurrentes (y los sitios están ligados) (Messer y Petrov 2013). Si hay mutaciones ligeramente deletéreas segregando en la población las cuales contribuyen al polimorfismo pero difícilmente se fijan, detectar la evolución adaptativa resultará más complicado porque siempre será una subestima,  $P_n/P_s \gg D_n/D_s$  (Fay *et al.* 2001). Hay muchas evidencias a favor de la presencia de alelos ligeramente deletéreos segregando en las poblaciones (sección 1.1.2). Tradicionalmente esto se ha corregido obviando los alelos a baja frecuencia (Fay *et al.* 2001, 2002; Bierne y Eyre-Walker 2004; Andolfatto 2005; Charlesworth y Eyre-Walker 2006), sin embargo, no existe un consenso sobre

la frecuencia umbral a partir de la que se debería descartar polimorfismo. Charlesworth y Eyre-Walker (2008) demostraron que no hay muchas diferencias en el  $\alpha$  estimado eliminando alelos a partir del 15%, es decir, el valor de  $\alpha$  eliminando alelos por debajo del 15% es muy parecido al valor de  $\alpha$  eliminando todos los alelos por debajo del 80% por ejemplo, aunque también demostraron que esta corrección heurística no ayuda a recuperar el valor verdadero de  $\alpha$  completamente. Messer y Petrov (2013) desarrollaron una sencilla extensión heurística del MKT que funciona relativamente bien en presencia de arrastres selectivos recurrentes y miles de sitios segregantes. Para ilustrar esta extensión, definamos  $\alpha(x)$  como una función que depende de la frecuencia de los alelos derivados:

$$\alpha(x) = 1 - \frac{D_s}{D_n} \frac{P_n(x)}{P_s(x)} \quad (1.8)$$

Donde  $P_n(x)$  y  $P_s(x)$  son el número de sitios segregantes no-sinónimos y sinónimos respectivamente, con una frecuencia derivada  $x$ . Como  $\alpha(x)$  depende sólo de la razón  $P_n(x)/P_s(x)$ , cualquier sesgo (demográfico o debido a arrastres selectivos) que afecte al espectro de frecuencias lo hará por igual a los sitios no-sinónimos y sinónimos, cancelándose el uno al otro. En otras palabras, esta extensión consiste en estimar  $\alpha$  en intervalos de frecuencia y ajustar luego una función, donde el valor de  $\alpha(x)$  cuando  $x \rightarrow 1$  corresponde a la estima de este nuevo estimador. La función utilizada es una exponencial con la forma  $\alpha(x) \approx a + b^{-cx}$ .

### 1.3.3 NUEVAS APROXIMACIONES A LAS PRUEBAS DE NEUTRALIDAD

En los últimos 10-15 años los esfuerzos estadísticos se han diversificado y ya no sólo se diseñan pruebas de neutralidad robustas a la demografía y a la selección ligada. Actualmente se diseñan modelos donde la selección (y otras fuerzas evolutivas) están parametrizadas y pueden inferirse a partir de los datos. Por lo tanto, estas estimas serán buenas tanto en cuanto el modelo asumido sea bueno.

Muchos modelos sencillos permiten estimar sus parámetros mediante métodos de máxima verosimilitud. Hay otros métodos no frecuentistas para inferir parámetros como la probabilidad *a posteriori* máxima (MAP) de la escuela bayesiana. La inferencia de parámetros mediante máxima verosimilitud es uno de los métodos de inferencia más comunes, sino el más común, en genética de poblaciones. Sin embargo, los modelos cada vez son más complicados (o completos) y estimar la verosimilitud (*likelihood*) de un modelo genético poblacional complejo es muchas veces imposible. Afortunadamente existen métodos para inferir parámetros no basados en la estima de la verosimilitud (*likelihood-free inference framework*) como la computación Bayesiana aproximada (ABC) (Beaumont *et al.* 2002) o algoritmos de aprendizaje automático (más conocidos como *machine learning algorithms*) (Jones 2014). Algunos de los trabajos más influyentes que han inferido parámetros de modelos genético poblacionales complejos mediante ABC son: Becquety Przeworski (2007); Jensen *et al.* (2008); Peter *et al.* (2012) y mediante algoritmos de aprendizaje automático son: Pavlidis *et al.* (2010; 2013); Ronen *et al.* (2013); Lin *et al.* (2011; 2013); Pybus *et al.* (2015). En esta tesis este tipo de métodos de inferencia no han sido utilizados, pero sí la inferencia mediante el método de máxima verosimilitud, véase ejemplo a continuación.

## ESTIMADORES DE LA *DFE* DE NUEVAS MUTACIONES DELETÉREAS Y LA TASA DE EVOLUCIÓN ADAPTATIVA

De todos los modelos genético poblacionales que utilizan la inferencia de parámetros por máxima verosimilitud he elegido como ejemplos representativos un grupo de modelos muy similares que estiman la *DFE* de nuevas mutaciones deletéreas y la tasa de evolución adaptativa entre especies (Boyko *et al.* 2008; Keightley y Eyre-Walker 2007; Eyre-Walker y Keightley 2009; Schneider *et al.* 2011). Todos estos métodos son extensiones del MKT donde se infiere la historia demográfica de la población a partir del espectro de frecuencias de sitios sinónimos (o neutros). A partir de esta demografía infieren por un lado la *DFE* de las nuevas mutaciones deletéreas (Boyko *et al.* 2008; Keightley y Eyre-Walker 2007; Eyre-Walker y Keightley 2009) y por otro lado la *DFE* de las nuevas mutaciones adaptativas (Schneider *et al.* 2011) que ocurren sobre sitios no-sinónimos (o funcionales), respectivamente. Los métodos de Boyko *et al.* (2008) y Eyre-Walker y Keightley (2009) aunque no estiman *per se* la *DFE* de nuevas mutaciones no-sinónimas adaptativas, sí son capaces de estimar el número de substituciones no-sinónimas adaptativas entre especies. Esto lo consiguen estimando primero el número de mutaciones no-sinónimas ligeramente deletéreas y neutras fijadas entre especies a partir de la *DFE* de las mutaciones deletéreas y substrayendo después este valor esperado al valor observado de substituciones no-sinónimas. En cambio, el método de Schneider *et al.* (2011) es capaz, no sólo de estimar el número de mutaciones adaptativas fijadas, sino de estimar también la tasa de mutación y el coeficiente de selección de las nuevas mutaciones adaptativas. *DFE-alpha* es el software creado por Peter Keightley (<http://www.homepages.ed.ac.uk/pkeightl/software>) el cual integra, el método para estimar la *DFE* de nuevas mutaciones deletéreas diseñado por Keightley y Eyre-Walker (2007), y el método para estimar las substituciones adaptativas de Eyre-Walker y Keightley (2009) y Schneider *et al.* (2011). En esta tesis hemos utilizado tanto los estimadores desarrollados por Keightley y Eyre-Walker (2007) y Eyre-Walker y Keightley (2009) y se explican en detalle a continuación bajo las siglas PKEW.

PKEW modela la *DFE* de las nuevas mutaciones deletéreas que ocurren en los sitios funcionales asumiendo una población diploide con generaciones discretas y panmixia. Esta población de tamaño  $N_1$  se asume que ya se encuentra en el equilibrio mutación-selección-deriva y puede experimentar (en una sola generación) un cambio en el tamaño poblacional (expansión o contracción) a un tamaño  $N_2$ . La población puede permanecer con este nuevo tamaño  $N_2$  a lo largo de  $t_2$  generaciones hasta el presente. Es en el presente cuando el espectro de frecuencias (SFS) para sititos funcionales y neutros es estimado en una muestra de  $n$  individuos. Todos los sitios se asume segregan independientemente y todos tienen la misma tasa de mutación, se asume también que la tasa de mutación es lo suficientemente baja como para que sólo dos alelos segreguen en un mismo sitio (*infinite-sites model*, Kimura 1969b; 1971; Watterson 1975). Existen dos tipos de sitios, unos sin efectos sobre la *fitness*, o neutros, y otros funcionales los cuales sí están sujetos a la aparición de nuevas mutaciones deletéreas (o neutras). La *fitness* del genotipo *wild-type*, heterocigoto y homocigoto para la nueva mutación es 1,  $1-s/2$  y  $1-s$ , respectivamente, es decir, son mutaciones con efecto aditivo. Cada mutación tiene un coeficiente de selección  $s$  que proviene de una distribución gamma definida por dos parámetros: la fuerza promedio de las nuevas mutaciones deletéreas,  $\gamma = -N_e s$ , y el parámetro de forma (*shape parameter*)  $\beta$ , esto permite que la distribución pueda adquirir un amplio abanico de formas, desde una distribución uniforme a una exponencial. El usuario sólo tiene que proporcionar el SFS de una región o genoma dado, el número de substituciones y el total de sitios sinónimos (o neutros) y no-sinónimos (o funcionales) y *DFE-alpha* se encarga de estimar los parámetros  $\gamma$ ,  $\beta$ ,  $N_2/N_1$ ,  $t_2$ , y el número de substituciones adaptativas. Es importante destacar que si  $N_2/N_1 = 1$  este modelo corresponde al modelo de Wright-Fisher con selección (sección 1.1.1).

No obstante, en la sección 1.1.3 y la sección 1.2.1 se dijo que la falta de independencia entre alelos provoca por un lado la variación del  $N_e$  (y la eficiencia de la selección) a lo largo del genoma (Hill y Robertson 1966) y por otro puede dificultar

la estimación de la demografía si no hay regiones en el genoma libres de la selección ligada (Hahn 2008; Comeron 2014), respectivamente. Dado que el método de PKEW asume que los sitios segregan independientemente, cabe preguntarse si este es un buen método para estimar la tasa de evolución y la demografía en genomas reales. Es decir, en genomas sujetos a cambios demográficos y con recombinación restringida entre sitios. Eyre-Walker y Keightley (2009) demostraron que cuando no hay cambios demográficos las estimas de la fracción de substituciones adaptativas ( $\alpha$ ) están prácticamente libres de sesgo tanto bajo segregación independiente como cuando la recombinación está muy restringida. Simulando genes de distintos tamaños sin recombinación, pero independientes entre sí, las estimas de  $\alpha$  obtenidas eran indistinguibles del valor verdadero (a no ser que la secuencia codificadora fuese de 50 kb o más longitud, en este caso  $\alpha$  se sobreestima). Cuando el constreñimiento en la fase divergente es mayor (o menor) al constreñimiento en la fase polimórfica, esto sucede cuando el  $N_e$  en la fase divergente es mayor (o menor) del  $N_e$  en la fase polimórfica (el cual abarca en promedio las últimas  $4N_e$  generaciones), la subestima (o sobreestima) de  $\alpha$  depende de la *DFE* y de la intensidad del cambio demográfico; aquellas *DFE* con una mayor proporción de mutaciones casi neutras ( $-1 < N_e s < 1$ ) son más susceptibles a este tipo de sesgos. En cualquier caso, como la demografía tiene efectos sobre todo el genoma, todo él se verá afectado por igual si los  $N_e$  a corto y largo plazo son muy distintos.

Messer y Petrov (2013) también respondieron a esta pregunta mediante simulación de cromosomas con características propias de humanos en lo que respecta a la tasa de mutación, censo efectivo, *DFE* para mutaciones deletéreas, densidad génica y tasa de recombinación. A partir de aquí jugaron con la fuerza de la selección positiva y el número de nuevas mutaciones beneficiosas. Cuando aplicaban el método de PKEW con la corrección demográfica observaban que el valor de  $\alpha$  estimado era prácticamente el mismo que el valor real, pero curiosamente estimaban siempre una expansión demográfica a pesar de que la población simulada había permanecido a

tamaño constante. Este resultado indica que la selección ligada afecta al SFS de los sitios neutros del mismo modo que una expansión demográfica reciente (como ya se comentó en la sección 1.2.1) y que al intentar corregir el efecto de la demografía sobre la *DFE* este método es capaz de corregir también el efecto Hill-Robertson sobre la *DFE*.

En resumen, modelos sencillos como este (donde se estiman tan solo 4 parámetros) todavía tienen mucho recorrido en genética de poblaciones pues son capaces de encapsular procesos evolutivos aparentemente muy diferentes (como el efecto Hill-Robertson, la selección ligada y la demografía) en una misma variable y ayudarnos a estimar cantidades importantes, como la tasa de evolución adaptativa entre especies, de forma “segura” a pesar de las perturbaciones impuestas por la demografía y la interferencia entre alelos.

## 1.4 *Drosophila melanogaster*

La mosca de la fruta *Drosophila melanogaster* es uno de los modelos experimentales más exitosos de la investigación genética (Roberts 2006). Se introdujo a principios del siglo XX como herramienta para el análisis genético y resultó ser crucial en los primeros pasos de la genética (Morgan 1915; Muller 1927) y también para la rama de la genética de poblaciones. Desde entonces se ha mantenido a la vanguardia de la investigación, proporcionando información empírica sobre los factores que afectan la variación genética en las poblaciones naturales (Ayala *et al.* 1974; Singh y Rhomberg 1987; Powell 1997), la construcción del plano corporal de los animales (Lewis 1978; Nüsslein-Volhard y Wieschaus 1980) y la función del sistema nervioso (Ivanov *et al.* 2004), entre otros. La gran cantidad de información funcional disponible (The modENCODE Project Consortium 2010; Gallo *et al.* 2011) y las múltiples técnicas de manipulación genética que existen, hace que *D. melanogaster* sea un organismo muy adecuado para el estudio de las fuerzas evolutivas que determinan la variación genética de las poblaciones naturales. Podemos afirmar por lo tanto que *Drosophila* ha sido y continúa siendo uno de los organismos modelo favoritos de la genética de poblaciones.

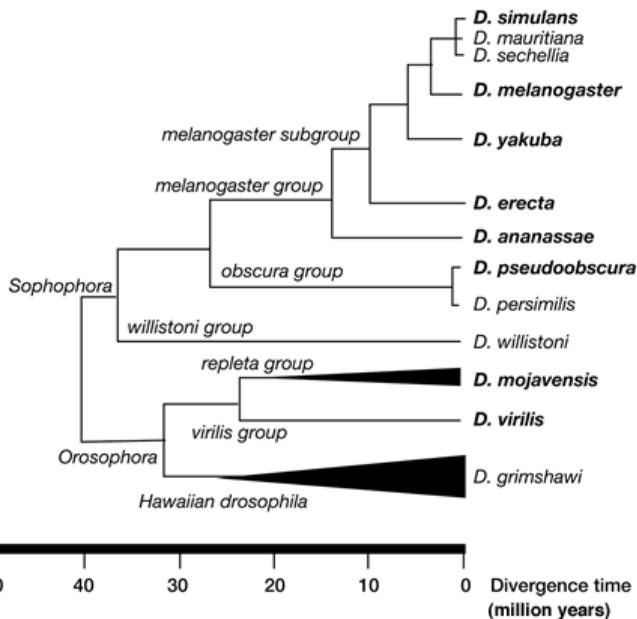
### CARACTERÍSTICAS DEL GENOMA

El genoma de *D. melanogaster* fue el tercero en ser secuenciado en eucariotas (Adams *et al.* 2000), después de *Saccharomyces cerevisiae* (Goffeau *et al.* 1996) y *Caenorhabditis elegans* (The *C. elegans* Sequencing Consortium 1998), y el primer organismo eucariota secuenciado mediante tecnología *shotgun* (Rubin 1996, Adams *et al.* 2000), un predecesor de las actuales tecnologías *next generation sequencing*. Desde entonces, se han generado una multitud de nuevas herramientas genómicas, incluyendo bases de datos especializadas y herramientas de análisis de secuencias (Matthews *et al.* 2005; Fox *et al.* 2006; Galperin 2007) que han permitido mejorar la secuencia genómica inicial en calidad y riqueza (Ashburner y Bergman 2005), en la

determinación de regiones ambiguas como los huecos o *gaps* (Celniker *et al.* 2002; Hoskins *et al.* 2002) así como en una definición extensiva de las anotaciones funcionales (Kopczynski *et al.* 1998; Berger *et al.* 2001; Kaminker *et al.* 2002, Misra *et al.* 2002; Carvalho *et al.* 2003; Bergman *et al.* 2005; Tupy *et al.* 2005; Hoskins *et al.* 2011; Hoskins *et al.* 2015). *D. melanogaster* sigue siendo hoy día uno de los genomas eucariotas mejor anotados. El genoma de *D. melanogaster* mide aproximadamente 180Mb repartidos en 5 cromosomas; X, Y, 2, 3 y el 4 o *dot* (Hoskins *et al.* 2002). Se han anotado 17.716 genes (versión 6.05 del genoma de referencia). Se estima que el genoma eucloromático (haploide) de *D. melanogaster* consta de 120Mb, de los cuales un 15% codifica para proteínas, un 3,2% es UTR, un 30% pertenece a secuencia intrónica y un 51,8% intergénica (Adams *et al.* 2000; Misra *et al.* 2002; Alexander *et al.* 2010).

## FILogenia e Historia Demográfica

Los artrópodos son uno de los taxones animales más diversos y exitosos, pues representan aproximadamente el 75% de todas las especies animales. En el linaje de los insectos se encuentra el taxón *Drosophila* que lo conforman más de 2.000 especies (Powell 1997). Estudios filogenéticos basados en la tasa de mutación genómica han determinado que los principales linajes dentro del género *Drosophila* divergieron hace unos 40 - 62 MYA (Russo *et al.* 1995; Tamura *et al.* 2004). Uno de los linajes lleva al subgénero *Sophophora*, formado por ~330 especies conocidas, entre las que se encuentra la especie *D. melanogaster* junto con especies como *D. simulans* o *D. yakuba* utilizadas habitualmente como taxones externos del organismo modelo. El otro linaje dio lugar a los subgéneros *Drosophila* e *Idiomyia* (*Drosophila*s hawaianas), los cuales están formados por ~1.100 y ~380 especies identificadas, respectivamente.



**FIGURA 1.10** Filogenia del género *Drosophila*. Árbol filogenético de 13 especies del género *Drosophila*. [Figura tomada de <http://eisenlab.org/AAA/melanogaster.html>].

*D. melanogaster* es una especie originaria del África subsahariana que se ha expandido recientemente al resto del mundo convirtiéndose así en una especie cosmopolita (Lachaise *et al.* 1988; David y Capy 1988; Begun y Aquadro 1993; Andolfatto 2001; Stephan y Li 2007). La expansión desde África hacia Europa se ha datado aproximadamente entre hace 10.000 y 19.000 años, lo que equivale a 0,1 - 0,4  $N_e$  generaciones (Li y Stephan 2006; Thornton y Andolfatto 2006; Duchen *et al.* 2013). Sin embargo, las primeras evidencias de que *D. melanogaster* llegó a América del Norte son de hace menos de 200 años (Johnson 1913; Sturtevant 1920; Keller 2007). Debido a que en aquella época la fauna díptera estaba muy bien descrita, es improbable que los entomólogos pasaran por alto a *D. melanogaster* durante muchos años (Keller 2007). En menos de 25 años, esta especie se propagó por todo el continente convirtiéndose así en uno de los dípteros más comunes en América del Norte (Howard 1900). La variación encontrada en cualquier población no africana es menor a la encontrada en África (Begun y Aquadro 1993; Andolfatto 2001) y parece, por tanto,

que *D. melanogaster* ha sufrido una reducción en su censo de población (Akashi 1996). Esto sugiere que la propagación fuera de África de las distintas poblaciones ha sido precedida por un cuello de botella (Begun y Aquadro 1993; Andolfatto 2001; Li y Stephan 2006; Thornton y Andolfatto 2006). Dicho cuello de botella se cree finalizó hace ~ 50 años o hace menos de 0,0042  $N_e$  generaciones (Thornton y Andolfatto 2006; Karasov *et al.* 2010).

Además, analizando la diferenciación de las poblaciones se puso de manifiesto que las poblaciones americanas eran genéticamente más cercanas a las poblaciones africanas que a las europeas (Caracristi y Schlötterer 2003; Baudry *et al.* 2004; Haddrill *et al.* 2005). Un estudio sugiere que la población norteamericana es una mezcla migratoria entre poblaciones y estiman que la proporción ancestral africana es del 15% y la europea del 85% (Duchen *et al.* 2013). Con esta distribución cosmopolita se espera que las diferentes poblaciones hayan evolucionado y adaptado de forma diferente a los distintos ambientes, haciendo de *Drosophila* un organismo modelo apropiado para el estudio de la adaptación y/o de la demografía poblacional; en definitiva, para la genética de poblaciones.

## GENÓMICA DE POBLACIONES

Hasta 2007, con el trabajo pionero de Begun *et al.* (2007) en *D. simulans*, los estudios de genética de poblaciones en *Drosophila* se han basado en muestras genómicas fragmentarias y no aleatorias del genoma, esto puede generar una visión parcial o sesgada de los procesos de genética de poblaciones. Aunque el estudio de Begun *et al.* (2007) puede considerarse el primero basado en un conjunto de datos genómico poblacional, seguido por el reanálisis de Macpherson *et al.* (2007) y los estudios genómico-poblacionales en levaduras (Liti *et al.* 2009) y *D. melanogaster* (Sackton *et al.* 2009), estos estudios estaban basados en secuenciación de baja cobertura con pocos individuos. Por ejemplo, el estudio en *D. simulans* (Begun *et al.* 2007;

Macpherson *et al.* 2007) y *D. melanogaster* (Sackton *et al.* 2009) se basaron en secuencias de baja cobertura con 3,9 lecturas por base en 7 líneas y 5,4 lecturas por base en 6 líneas, respectivamente. Valores manifiestamente insuficientes para todo estudio de genética de poblaciones cuyas inferencias se basen en la frecuencia de las variantes. Se ha estimado que para obtener buenas secuencias son necesarias coberturas mínimas de 10-15X (Craig *et al.* 2008, Smith *et al.* 2008). A pesar de estas limitaciones técnicas los resultados de Begun *et al.* (2007) han desafiado muchas de las predicciones esperadas bajo la teoría neutra (Kimura 1971) (véase sección 1.2.1) y ha abierto el camino a nuevos estudios de genómica de poblaciones más masivos.

Dicho esto, el *data desideratum* para un estudio completo y no sesgado de variación genética consiste en una gran muestra de genomas completos de alta calidad de una misma población natural de un organismo del que exista un extenso conocimiento biológico y la secuencia completa, también de calidad, de un taxón externo con tal de estimar la divergencia. Esto permite una descripción completa y detallada de los patrones de variación a gran escala. *D. melanogaster* es una de las especies de la que más datos genómico poblacionales disponemos, actualmente hay 623 genomas completos secuenciados a alta cobertura. El *Drosophila Genome Nexus* (Lack *et al.* 2015) es un recurso genómico poblacional con 623 genomas completos secuenciados a alta cobertura el cual ha reensamblado los 4 grandes conjuntos de datos genómico poblaciones disponibles (Langley *et al.* 2012; Mackay *et al.* 2012; Pool *et al.* 2012; Huang *et al.* 2014) y ha añadido más genomas de poblaciones sub-Saharianas. Este reensamblaje es necesario si pretendemos analizar el conjunto completo de genomas. Uno de estos estudios es el *Drosophila Genetic Reference Panel* (DGRP) (Mackay *et al.* 2012) (véase sección 2.1 para más detalles) al cual tuvimos acceso antes que el resto de la comunidad científica, pues nuestro grupo dirigido por el Dr. Barbadilla forma parte del consorcio internacional encargado de los análisis de genética de poblaciones utilizando SNPs. Los resultados genético poblacionales más relevantes de Mackay *et al.* (2012) basados en 158 genomas completos secuenciados

con una alta cobertura de una población norteamericana son: (1) la falta de correlación entre polimorfismo y recombinación por encima de 2 cM/Mb (véase Barrón 2015 para más detalles), (2) la estima de la proporción del genoma que está sometida a la acción de la selección purificadora (~40%), (3) la estima de la fracción de substituciones adaptativas a nivel genómico diferenciando entre clases de sitios codificadores y no-codificadores y contextos cromosómicos – globalmente se estima que ~ 25% de las sustituciones son adaptativas y que el tercio centromérico de los autosomas muestran escasas evidencias de selección positiva y (4) el desarrollo de un visualizador de la variación genética (<http://popdrowser.uab.cat/>) (Ràmia *et al.* 2012) que integra todos los estadísticos genético poblacionales estimados en ventanas cromosómicas y a nivel de genes codificadores. Esta tesis es un estudio independiente al de Mackay *et al.* (2012) a pesar que se utiliza el mismo conjunto de datos genómico poblacional y algunos de los estimadores que se utilizaron y publicaron por primera vez en Mackay *et al.* (2012) han sido caracterizados mediante simulaciones y re-utilizados en el presente trabajo.

## 1.5 OBJETIVOS

Esta tesis es un proyecto de genómica de poblaciones cuyo objetivo global es describir y cuantificar la selección natural y el efecto Hill-Robertson a lo largo del genoma de *D. melanogaster*. Para ello se han seguido aproximaciones bioinformáticas y teórico-estadísticas. Los objetivos específicos de este proyecto son: (i) desarrollar dos nuevos estimadores puntuales de la acción de la selección purificadora basados en el contraste del espectro de frecuencias de sitios putativamente seleccionados y neutros, (ii) estimar la *DFE* de nuevas mutaciones deletéreas (sección 1.1.2) y la tasa de evolución adaptativa (sección 1.3) a lo largo del genoma codificador y no-codificador de *D. melanogaster*, (iii) cuantificar el papel del efecto Hill-Robertson (sección 1.1.3) sobre la adaptación proteica en *D. melanogaster*.

### i. DESARROLLAR ESTIMADORES DE LA ACCIÓN DE LA SELECCIÓN PURIFICADORA.

La evidencia en numerosas especies de la omnipresencia de la selección purificadora y sus efectos sobre la variación neutra ligada (Charlesworth 1993), requiere del desarrollo de pruebas estadísticas que sean capaces de medir la intensidad de dicha selección y a su vez que sean robustas a cambios demográficos recientes (véase sección 1.2.1). En este trabajo se proponen dos nuevos estimadores de la selección purificadora que se probarán con datos obtenidos mediante ecuaciones de difusión donde se evaluarán distintas *DFEs* para nuevas mutaciones deletéreas y su poder estadístico simulando un número limitado de sitios segregantes. Mediante simulaciones *forward in time* se obtendrán datos fuera del equilibrio, como los obtenidos tras un cuello de botella reciente y/o en presencia de selección de fondo. En concreto, se propondrán y pondrán a prueba bajo condiciones genómicas realistas dos estadísticos intuitivos basados en el contraste de distintas partes del espectro de frecuencias de sitios putativamente seleccionados y neutros.

- ii. **ESTIMAR LA DFE DE NUEVAS MUTACIONES DELETÉREAS Y LA TASA DE EVOLUCIÓN ADAPTATIVA EN EL GENOMA CODIFICADOR Y NO-CODIFICADOR DE *D. melanogaster*.** Gracias a la disponibilidad de 158 genomas completos secuenciados con una alta cobertura provenientes de una población natural de *D. melanogaster* (Mackay *et al.* 2012) se pretende obtener las estimas más completas y precisas de la *DFE* de nuevas mutaciones deletéreas y la tasa de adaptación hasta la fecha. Se compararán las estimas de nuestros nuevos estimadores de la selección purificadora con las estimas de la *DFE* obtenidas mediante el método de Keightley y Eyre-Walker (2007). Disponer de estas estimas nos ayudará a responder a estas cuestiones: (1) ¿Qué proporción del genoma es funcional? (2) ¿Cuál es la contribución relativa de la variación genética en secuencias codificadoras y no-codificadoras a la variación en la eficacia biológica? (3) ¿Cuál es la contribución relativa de las mutaciones fijadas en el genoma codificador y no-codificador a la evolución adaptativa? (4) ¿Existen diferencias en la *DFE* y la tasa de evolución adaptativa entre brazos cromosómicos? Si es así, ¿a qué se deben?
- iii. **CUANTIFICAR EL PAPEL DEL EFECTO HILL-ROBERTSON SOBRE LA ADAPTACIÓN PROTEICA EN *D. melanogaster*.** En este último objetivo se pretende correlacionar la tasa de recombinación, la densidad génica y la tasa de mutación con la tasa de adaptación proteica. Esto nos ayudará a comprender la importancia relativa de la interferencia Hill-Robertson en el genoma codificador de *D. melanogaster* y a su vez estimar cuantas substituciones adaptativas dejan de fijarse debido a la interferencia entre mutaciones seleccionadas (Barton 1995) (véase sección 1.1.3). Además, conocer la relación entre la tasa de adaptación y la tasa de mutación es de vital importancia - algunos trabajos sugieren que son independientes entre sí en *D. melanogaster* (véase cuadro 4).



# MATERIALES Y MÉTODOS

---

## 2.1 SECUENCIAS POBLACIONALES GENÓMICAS

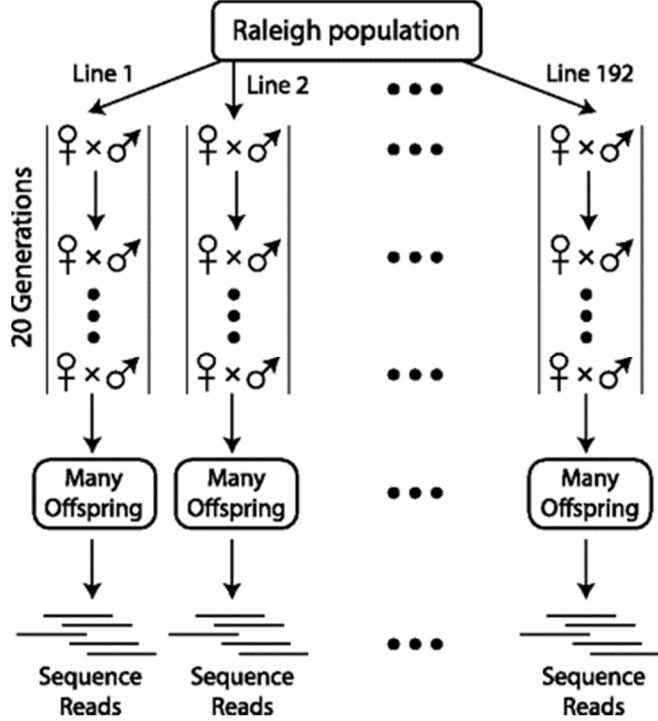
Esta tesis se ha realizado íntegramente analizando secuencias genómicas de la especie modelo *Drosophila melanogaster*. Las secuencias se obtuvieron en el proyecto DGRP (Mackay *et al.* 2012) a partir de una muestra de moscas de una población de Raleigh (Carolina del Norte, EEUU). Se analizaron los patrones de variación intra e interespecíficos de las variantes de un único nucleótido (SNPs). Algunos análisis complementarios han requerido también de los datos del proyecto DPGP2 (Pool *et al.* 2012) de una población de Gikongoro (Ruanda). Es importante destacar que el procesamiento y filtrado de los alineamientos genómicos y el cálculo de estadísticos de la población norteamericana fueron realizados por nuestro grupo, que participó en el consorcio DGRP. Los datos de la población africana fueron, en cambio, procesados y filtrados por el Dr. José L Campos de la *University of Edinburgh* y los detalles sobre la obtención de los datos y cálculo de estadísticos en la población africana se encuentran en los artículos originales (véase Campos *et al.* 2014).

### 2.1.1 SECUENCIAS PROYECTO DGRP

Los datos para el estudio de genómica de poblaciones provienen del proyecto internacional *Drosophila Genetic Reference Panel* (DGRP) (Freeze 1; Mackay *et al.* 2012). Este recurso está compuesto por un conjunto de 168 líneas consanguíneas individuales secuenciadas con una elevada cobertura (~22X en promedio y longitud promedio de las lecturas es de ~80 pb, consultese material suplementario en Mackay *et al.* (2012) sobre los detalles de la secuenciación, ensamblaje y el *calling* de los SNPs). El número de cromosomas original del proyecto DGRP pasó de 168 a 158 porque se detectaron distintos errores en 10 líneas consanguíneas (contaminaciones,

duplicaciones) – estas fueron eliminadas en el presente análisis. Los datos de variación poblacional junto a los genomas secuenciados de las especies cercanas a *D. melanogaster* (*D. simulans* y *D. yakuba* principalmente), han permitido aplicar una batería de pruebas de neutralidad (véase sección 1.3.2 de la Introducción) así como sus extensiones basadas en la inferencia de parámetros por máxima verosimilitud (véase sección 1.3.3). Además, el hecho que se trate de líneas consanguíneas reduce en gran parte el problema de la inferencia de la fase haplotípica de las variantes genéticas en especies diploides.

Brevemente, las líneas se obtuvieron tras 20 generaciones de cruces endogámicos hermano-hermana a partir de una muestra de hembras grávidas recolectadas de una población natural de Raleigh (Carolina del Norte, EEUU) (figura 2.1). El coeficiente de consanguinidad (*F*) esperado después de las 20 generaciones es de 0,986 (Falconer y Mackay 1996), es decir, teóricamente casi el 99% de los heterocigotos se habrán fijado en una u otra línea, y sólo 1,4% de SNPs se espera segreguen como heterocigotos residuales en cada isolínea. No obstante, se observó una mayor proporción de heterocigotos residuales de lo esperado en algunas líneas. Trabajos posteriores (Huang *et al.* 2014) mostraron que el 92% de los brazos autosómicos de las isolíneas tenían entre 0,5-2% de heterocigotos residuales, aproximadamente lo esperado teóricamente, mientras que un 8% contenían 9% o más de heterocigotos residuales. De este 8% el 96% contenían al menos una inversión heterocigótica y el 4% restante el cariotipo estándar. Se demuestra, por tanto, que hay una correlación casi perfecta entre cromosomas con heterocigotos residuales y ser portador de inversiones heterocigóticas.



**FIGURA 2.1** Diseño experimental seguido para la obtención y secuenciación de las líneas DGRP. Cada línea DGRP fue fundada por una hembra grávida capturada en las afueras de Raleigh, Carolina del Norte. Cada nueva generación resulta del cruce de una pareja macho y hembra de la generación anterior. Cada línea DGRP se obtuvo después de 20 generaciones hermano-hermana. De cada línea se extrajo ADN de un acervo de 500-1000 moscas que fueron secuenciadas. [Figura tomada de Stone (2012)].

### 2.1.2 LÍNEAS CONSANGUÍNEAS Y VARIACIÓN NATURAL

Una cuestión recurrente es hasta qué punto la variación encontrada en estas líneas consanguíneas es representativa de la variación encontrada en la población natural de las cuales fueron derivadas. Esta es una cuestión crucial en nuestra inferencia de la variación natural, pues podría introducir sesgos importantes en nuestros estadísticos genético poblacionales y las conclusiones que derivemos de ellos. Primero de todo destacar que el número de líneas consanguíneas (en este caso 158) no equivale al número de moscas capturadas en su día. Es decir, cada línea consanguínea es “idealmente” una mosca haploide. Digo idealmente porque ya se ha mostrado anteriormente que se espera cierta proporción de heterocigotos residuales.

Por lo tanto, el recurso DGRP no representa a 158 moscas diploides, sino al espectro de frecuencias observados de 158 gametos o individuos haploides. Dicho esto, en la generación de las líneas consanguíneas tres procesos pueden alterar, potencialmente, el espectro de frecuencias: (1) eliminación de alelos recesivos deletéreos, (2) cambio en las presiones selectivas a las que están sometidos algunos alelos durante la generación de las líneas e (3) incorporación de nuevas mutaciones durante el proceso de generación de las líneas consanguíneas.

Los dos primeros efectos pueden argumentarse verbalmente de la siguiente forma: cada cruce hermano-hermana supone un cuello de botella de proporciones máximas, si este proceso de muestreo al azar de los cromosomas se repite 20 veces, es de esperar que la intensidad de la selección necesaria para contrarrestar y sesgar el espectro de frecuencias obtenido tras generar las líneas consanguíneas respecto al espectro esperado en un muestreo directo de la población natural debe ser muy alto (Spiess 1989; García-Dorado 2012). Analíticamente la expresión que define la condición en la que la selección y la deriva tienen la misma influencia sobre las frecuencias alélicas es  $2s = 1/2N_e$  (Crow y Kimura 1970; Lynch 2007), donde  $s$  es el coeficiente de selección absoluto. Si consideramos que en nuestro caso  $2N_e = 4$  (porque tenemos dos moscas), deducimos que sólo aquellos alelos con valores de  $|s| >> 1/8$  se eliminarán (o fijarán) por la fuerza de la selección durante el proceso de generación de las líneas consanguíneas. Del mismo modo, todos aquellos alelos con valores de  $|s| < 1/8$  se fijarán o extinguirán aleatoriamente al cabo de 20 generaciones con una probabilidad de  $p$  y  $1-p$ , respectivamente, donde  $p$  es la frecuencia del alelo en la muestra de 158 cromosomas de la población natural original. En otras palabras, la variación genética almacenada en las líneas consanguíneas puede considerarse una muestra representativa de la variación genética presente en la población natural en el momento del muestreo, siempre y cuando la mayoría de los alelos no tengan  $|s| >> 1/8$  o su coeficiente de selección en laboratorio haya aumentado a  $|s| >> 1/8$ . En cualquier caso, la probabilidad de

muestrear un alelo recesivo deletéreo con  $s = 1/8$  en la naturaleza es muy baja, ya que, la frecuencia esperada bajo el equilibrio mutación-selección es de  $q = \sqrt{u/s}$ , donde  $u$  es la tasa de mutación y  $s$  es el coeficiente de selección. Teniendo en cuenta que la tasa de mutación de *Drosophila* es del orden de  $10^{-9}$  por sitio y por generación (Sharp y Li 1989; McVean y Vieira 2001; Keightley *et al.* 2009; Schrider *et al.* 2013),  $q \sim 0,017\%$ , la probabilidad de muestrear una mutación de esta frecuencia en una muestra de 158 cromosomas es muy baja. También existe la posibilidad de incorporar nuevas mutaciones durante el proceso de generación de las líneas consanguíneas a lo largo de las 20 generaciones. Dado que en la población hay 4 copias de cada cromosoma, el genoma eucromático (haploide) de *D. melanogaster* mide  $\sim 120$  MB, la tasa de mutaciones puntuales es de  $10^{-9}$  por sitio y por generación, y han ocurrido 20 generaciones, se habrán generado  $\sim 10$  mutaciones *de novo* por línea consanguínea. No obstante, no todas estas mutaciones se fijarán, como cada nueva mutación aparece en 1 de cada 4 cromosomas presentes en la población (hermano-hermana), esperamos que cada línea consanguínea fije  $1/4$  de las 10 mutaciones que han ocurrido durante la generación de la línea consanguínea. Luego, sólo  $\sim 2,5$  variantes por línea será variación no natural y todas estarán en copia única (*singletons*). Si tenemos 158 líneas en las que este proceso puede ocurrir, habrá un promedio de 420 SNPs *de novo* de los aproximadamente 4.800.000 SNPs descubiertos en las secuencias. Esto supone un sesgo potencial insignificante.

En conclusión, *a priori* la metodología seguida para generar las líneas consanguíneas no parece que pueda generar grandes sesgos en los patrones de variación nucleotídica observados en las líneas respecto a los patrones que esperaríamos encontrar en la población natural.

### 2.1.3 CUELLOS DE BOTELLA, ESTRUCTURA POBLACIONAL Y VARIACIÓN NATURAL

Otra importante peculiaridad de nuestra población es su historia demográfica, la cual está lejos de representar (como en casi todas las poblaciones naturales) una población panmíctica con un censo efectivo constante. Un cuello de botella severo y reciente, como el experimentado por nuestra población norteamericana (Li y Stephan 2006; Thornton y Andolfatto 2006; Duchen *et al.* 2013) conduce a la rápida extinción o fijación de los alelos derivados y por lo tanto a un defecto de alelos a baja frecuencia respecto a lo que esperaríamos en una población de tamaño constante (Allendorf 1986; Denniston 1978; Nei *et al.* 1976; Maruyama y Fuerst 1985; Watterson 1984; Luikart *et al.* 1998). Esto conlleva a una reducción global de los niveles de variación genética: en África la variación nucleotídica putativamente neutra es el doble (Andolfatto 2005; Singh *et al.* 2007; Pool *et al.* 2012) que en Norteamérica (donde  $n \sim 0,01$ ) (Singh *et al.* 2007; Sackton *et al.* 2009; Mackay *et al.* 2012; Langley *et al.* 2012; Pool *et al.* 2012).

No tener en cuenta la demografía de las especies puede conducir a inferencias incorrectas de las estimas de los parámetros que describen los procesos genético-poblacionales. Por esta razón, el efecto de la demografía sobre los nuevos estadísticos que proponemos para estimar la fuerza y eficacia de la selección purificadora han sido probados en la sección de resultados 3.1. A su vez, los estimadores que utilizamos para estimar la *DFE* (Keightley y Eyre-Walker 2007) y la tasa de evolución adaptativa (Eyre-Walker y Keightley 2009) han sido diseñados precisamente para corregir el efecto de la demografía reciente sobre las estimas (véase sección 1.3.3). Es decir, a pesar que la población no cumple algunos de los supuestos en los que están basados muchos trabajos teóricos clásicos (como un censo efectivo constante), los estadísticos y métodos de genética de poblaciones utilizados en esta tesis se ponen a prueba, o se han puesto a prueba previamente por otros autores, en presencia de cambios demográficos recientes.

Finalmente, la estructura poblacional puede ser otra fuente importante de error a la hora de interpretar los parámetros estimados tanto para estudios de genética de poblaciones como de genética cuantitativa. Por ello se probó si había estructura poblacional, mediante un análisis filogenético de todas las isolíneas DGRP. El resultado fue un árbol en forma de estrella en el que cada una de las líneas se sitúa a una distancia filogenética similar a cualquier otra (material suplementario en Mackay *et al.* 2012). Además, también se realizó un análisis de componentes principales (PCA, *Principal Component Analysis*) de la matriz de covarianza genética (véase tabla suplementaria 2 en Mackay *et al.* 2012). Ambos análisis indicaron la práctica ausencia de estructura poblacional en las líneas DGRP.

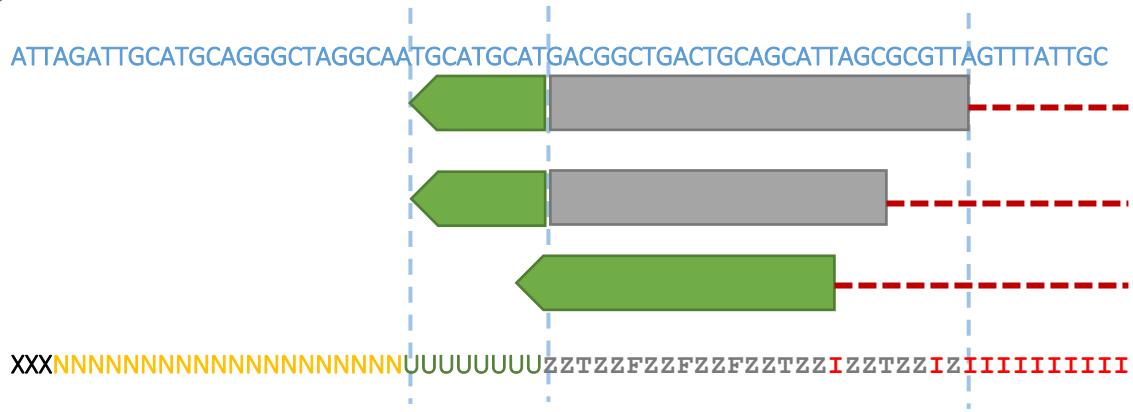
## 2.2 TRATAMIENTO DE LOS DATOS Y CÁLCULO DE LOS ESTADÍSTICOS

Este estudio se ha llevado a cabo en los brazos autosómicos (2L, 2R, 3L y 3R) y el cromosoma X de *D. melanogaster* utilizando como genoma de referencia la versión 5 del *Berkeley Drosophila Genome Project* (BDGP 5, <http://www.fruitfly.org/sequence/release5genomic.shtml>).

### 2.2.1 INCORPORACIÓN DE LAS ANOTACIONES GÉNICAS A LOS DATOS GENÓMICO POBLACIONALES

Conocer la clase funcional de cada base del genoma no es trivial incluso en uno de los genomas mejor anotados como es el de *Drosophila*. En algunos casos una misma base o posición del genoma tiene distintas funciones anotadas. Por ejemplo, una posición puede actuar como UTR para un transcripto y como 0-veces degenerada para otro transcripto. A su vez hay intrones con otros genes anidados dentro. ¿Cómo definimos en este último caso la región intergénica del gen que yace dentro de un intrón? Estos son ejemplos que ponen de manifiesto que cualquier estudio genómico que trate de clasificar las posiciones de acuerdo a las anotaciones disponibles tendrá

que establecer un criterio de análisis previo. En este estudio hemos clasificado cada posición del genoma de acuerdo a un criterio jerárquico. Si una posición está representada por más de una clase funcional esta ha sido “re-anotada” siguiendo este orden: 0-veces degenerado > 2-veces degenerado > UTR > Intrón > Intergénico > 4-veces degenerado.



**FIGURA 2.2** Ejemplo de secuencia “re-anotada” para un fragmento del genoma donde múltiples transcritos solapan. X es en negro la etiqueta para la secuencia intergénica que está a más de 5 kb del exón más cercano; N es en amarillo la etiqueta para secuencia intergénica a menos de 5 kb del exón más cercano, U es en verde la etiqueta para las secuencias UTR; Z, T y F son en gris las etiquetas para las posiciones cero, dos y cuatro veces degeneradas del código genético, respectivamente, e I es la etiqueta para intrones en rojo. La primera línea corresponde a la secuencia de ADN, la última línea en mayúsculas corresponde a la secuencia re-anotada consenso para un gen ortólogo 1:1 entre *D. melanogaster*-*D. yakuba*.

Atendiendo a este criterio se creó una nueva secuencia (que denominamos *re-anotada*, figura 2.2) que etiquetaba cada posición del genoma de acuerdo a la clase funcional a la que pertenecía. Además, si dicha secuencia está en mayúsculas implica que el gen es ortólogo uno a uno entre *D. melanogaster* y *D. yakuba*, si por el contrario está en minúscula el gen codificador y su respectiva secuencia no-codificadora se considera no ortóloga uno a uno y por lo tanto no ha sido analizada en este estudio. La versión del archivo de anotaciones utilizado en todo este trabajo es la 5.50 (<http://flybase.org/>, último acceso en marzo de 2013).

## 2.2.2 CRITERIOS APLICADOS PARA EL FILTRADO DE LOS DATOS

Para los análisis basados en ventanas no solapantes de 1 Mb, disponer de dicha secuencia re-anotada facilita mucho el cálculo de los distintos estadísticos genético poblacionales (véase siguiente sección), pues simplemente analizamos las bases etiquetadas para la clase funcional en la que estamos interesados y que están en mayúsculas. No obstante, debido a nuestro criterio de re-anotación del genoma, se debe destacar que no disponemos de las regiones intergénicas de aquellos genes que yacen dentro de intrones, pues estas aparecen anotadas como secuencia intrónica. Por esta razón, este tipo de secuencias intrónicas que podrían confundirse funcionalmente con secuencias intergénicas, y viceversa, no han sido analizadas. En otras palabras, las secuencias intrónicas corresponden a las secuencias de aquellos intrones sin otros genes anidados dentro y las secuencias intergénicas en ningún caso funcionan como intrones. Además, tanto las regiones 5' como 3' de las regiones intergénicas y UTRs han sido agrupadas bajo una misma etiqueta. Es decir, no se ha distinguido entre las regiones aguas arriba y aguas abajo de los genes. Finalmente, dentro de las regiones intergénicas sólo se han analizado las bases 5 kb adyacentes (como máximo) a la última posición y la primera posición del último y primer exón de un mismo gen codificador, respectivamente. En cambio, para los intrones no se ha aplicado ningún límite de longitud. En total el ~60% del genoma eucromático de *D. melanogaster* ha sido re-anotado y ha pasado a la siguiente etapa del análisis que corresponde al cálculo del SFS y la estimación de la divergencia (véase siguiente sección). El resto (~40%), o bien corresponde a intrones con genes codificadores dentro (de acuerdo a nuestros cálculos y partiendo de las anotaciones génicas de *Drosophila* alrededor de un 25% de los genes codificadores están en intrones), o a secuencias intergénicas que están más allá de 5 kb del exón más cercano, o bien no son ortólogos uno a uno entre *D. melanogaster* y *D. yakuba*.

Para los análisis basados en genes codificadores autosómicos, la secuencia codificadora de un gen está representada por la secuencia codificadora de todos sus

transcritos, no por la de su transcripto más largo. De este modo nos aseguramos que todas las secuencias codificadoras de un gen están presentes en nuestro análisis – independientemente de si provienen de exones constitutivos o alternativos. No obstante, en el genoma de *Drosophila* hay exones que solapan parcialmente, en este caso la secuencia que solapa sólo ha sido analizada para el exón más largo (en ningún caso hemos analizado la misma posición dos o más veces). Sólo aquellos genes sin mutaciones polimórficas de desplazamiento en el marco de lectura o codones *stop* tempranos fueron analizados. Sólo aquellos genes ortólogos uno a uno entre *D. melanogaster* y *D. yakuba* y sin *gaps* entre ambas secuencias de referencia fueron analizados. La ausencia de *gaps* es importante para evitar que posibles errores de alineamiento creen una correlación positiva artificial entre nuestras estimas de la tasa de mutación (véase sección 2.6) y la tasa de adaptación (véase sección 2.7). Además, con tal de evitar que potenciales sitios reguladores del *splicing* solapen con la secuencia codificadora, los 23 codones adyacentes a los intrones fueron descartados tal y como se describe en Warnecke y Hurst (2007). Este filtro es importante para evitar que genes con intrones tengan estimas inferiores de variación genética simplemente debido a la regulación del *splicing*. Por último, este estudio se realizó sólo con los genes autosómicos porque de haber combinado genes del cromosoma X con genes autosómicos habríamos introducido un sesgo importante en nuestro conjunto de datos. Los genes del cromosoma X tienen un patrón de variación diferenciado del de los genes autosómicos (véase la sección de discusión 4.2.2). Además, el cromosoma X no tiene genes suficientes (menos de 900 que satisfagan nuestros filtros) como para realizar un estudio independiente del mismo calibre que el realizado para los genes autosómicos. En definitiva, el conjunto de datos final que satisface todos nuestros criterios de calidad y filtros es de 6.141 genes codificadores. Un conjunto de datos menor con 3.369 genes codificadores se ha utilizado también. En este conjunto de datos los intrones cortos (menores a 66 pb) fueron utilizados como alternativa a las posiciones 4-veces degeneradas para estimar la tasa de evolución adaptativa (véase sección 2.7). Siguiendo el trabajo de Halligan y Keightley

(2006), las posiciones 8-30 de intrones < 66 pb se utilizaron como secuencia de referencia neutra. Sólo aquellos genes con dos de estos intrones y un porcentaje de *gaps* (en la secuencia intrónica) menor al 10% entre *D. melanogaster* y *D. yakuba* fueron analizados.

### 2.2.3 ESTIMACIÓN DEL ESPECTRO DE FRECUENCIAS (SFS) Y LA DIVERGENCIA

Para estimar el espectro de frecuencias para cada una de las clases funcionales excluimos aquellas posiciones del genoma que mostraban un exceso de heterocigotos residuales o un elevado porcentaje de bases sin especificar (etiquetadas como *N* en los archivos *fasta*). Siendo más precisos, una posición del genoma ha sido descartada cuando menos de 128 cromosomas han podido ser analizados. De este modo conseguimos analizar el 97.8% del genoma eucromático re-anotado siguiendo los pasos descritos anteriormente. Hemos reducido el número de cromosomas analizados porque el método utilizado para estimar la distribución de coeficientes de selección de las nuevas mutaciones deletéreas y la tasa de adaptación (véase sección 2.7), así como nuestros nuevos estimadores de la eficacia y fuerza de la selección purificadora (véase sección 3.1), requieren que todas las posiciones estén representadas por el mismo tamaño de muestra. La reducción del número de cromosomas inicial de 158 a 128 se realizó muestreando al azar sin reemplazamiento cada posición del genoma (después de descartar los heterocigotos residuales y *N*). En definitiva, se utilizaron 128 cromosomas para estimar el *minor allele frequency* (MAF), o *folded site frequency spectrum* (folded-SFS).

Para estimar la divergencia se cogió un cromosoma al azar de la población norteamericana de *D. melanogaster* y se comparó como taxón externo con el genoma de referencia de *D. yakuba*. Se ha utilizado esta especie porque su genoma presentaba una de las coberturas más altas (9,1X) entre los genomas secuenciados por el consorcio de los 12 genomas del género *Drosophila* (*Drosophila 12 Genome Consortium*,

2007) y grupo de especies filogenéticamente más cercanas a *D. melanogaster*. El alineamiento entre las secuencias del proyecto DGRP, el genoma de referencia de *D. melanogaster* y el genoma de referencia de *D. yakuba* es público y se encuentra en <http://popdrowser.uab.cat/> (Ràmia *et al.* 2012). Además de su mayor cobertura, el uso de la secuencia de *D. yakuba* es más adecuado para obtener estimas de adaptación más precisas gracias al tiempo de divergencia óptimo transcurrido entre las dos especies (Keightley y Eyre-Walker 2012). Si la variación genética entre especies es baja respecto a la variación dentro de la especie (como es el caso de *D. melanogaster* y *D. simulans*), las estimas de la tasa de evolución adaptativa y la DFE pueden estar sesgadas debido a la contribución de polimorfismo ancestral a la divergencia. Los autores proponen que la divergencia estimada a partir de especies estrechamente emparentadas (como *D. melanogaster* y *D. simulans*, o humanos y chimpancés) puede inducir a subestimar la tasa de evolución adaptativa en un ~10% o más. Por tanto, utilizar *D. yakuba* permite obtener estimas más fiables de la selección adaptativa.

En este trabajo se han contado el número de los distintos sitios codificadores y no-codificadores basándose en la posición que ocupa en la secuencia. Centrándonos en las secuencias codificadoras, hemos contado el número de sitios, polimorfismos y el número de substituciones sinónimas y no-sinónimas para los sitios 0-veces (no-sinónimos) y 4-veces degenerados (sinónimos) por separado – los sitios 2 y 3 veces degenerados fueron descartados. No obstante, el hecho de haber definido los sitios según su posición física implica restringir nuestro análisis a aquellos tripletes que codifican el mismo aminoácido en las dos especies (*D. melanogaster* – *D. yakuba*). Al restringir nuestro análisis a aquellos codones que no muestran cambios no-sinónimos asumimos que ese codón no ha sufrido ninguna substitución de cambio de aminoácido. Esto nos ahorra tener que computar las distintas rutas mutacionales que han podido ocurrir entre dos codones que difieren por más de un cambio y es un supuesto razonable dada la baja divergencia no-sinónica entre las dos especies.

Los métodos más sofisticados que existen para estimar la tasa de sustitución sinónima y no-sinónima (como el de Li 1993 y Goldman y Yang 1994) definen los sitios, en cambio, como “oportunidades-mutacionales” – el número de sitios sinónimos equivale al potencial número de mutaciones o cambios sinónimos. Lo mismo ocurre para los sitios no-sinónimos. La razón principal para definir los sitios por posición física fue que la definición de los sitios como oportunidades-mutacionales puede dar una visión incorrecta de la correlación entre el sesgo en el uso de codones y la tasa de sustitución sinónima, ya que, para estos métodos el número de sitios sinónimos (y con ello la tasa de substitución por sitio) depende del propio sesgo en el uso de codones (consúltese Bierne y Eyre-Walker 2003). Como en este trabajo se ha estudiado por separado el efecto de ambas variables; sesgo en el uso de codones y tasa de sustitución sinónima, sobre la tasa de adaptación no-sinónima (véase sección 3.3), utilizar un estimador de la tasa de substitución sinónima que corrija para las diferencias en el sesgo en el uso de codones entre genes carece de sentido.

Una vez hemos contado los sitios pertenecientes a cada clase funcional codificadora y no-codificadora, en este caso siguiendo un criterio posicional, el número de cambios observados en la secuencia entre *D. melanogaster* y *D. yakuba* debe corregirse para la posibilidad de que múltiples mutaciones o cambios se hayan fijado en una misma posición desde la separación de ambas especies (*multiple-hits correction*). Existen multitud de métodos que permiten hacer esto, el método de Jukes y Cantor (1969) y Tamura (1992) fueron los elegidos en nuestro caso para corregir las estimas de divergencia en todas las clases de sitios. El método de Tamura (1992) permite dar estimas más precisas del número de substituciones cuando el contenido *GC* es variable y hay un sesgo a favor de las transiciones (*ts*) en lugar de las transversiones (*tv*) (situaciones que el método de Jukes y Cantor no considera). Es decir, el método de Tamura (1992) captura la característica principal del sesgo en el uso de codones en el genoma nuclear de *Drosophila*, que consiste en un enriquecimiento en *GC* en la última posición de los codones (Moriyama y Hartl 1993; Akashi 1994; Duret y

Mouchiroud 1999). Por esta razón las estimas de divergencia para las posiciones 4- veces degeneradas que se muestran en esta tesis son las corregidas utilizando el método de Tamura (1992). El método de Jukes y Cantor (1969) ha sido el elegido para corregir para múltiples cambios en el resto de clases funcionales ya que no hay grandes diferencias entre ambas metodologías para este tipo de sitios.

## 2.3 SIMULACIONES DE LOS ESTIMADORES DE LA SELECCIÓN PURIFICADORA

Con tal de investigar en qué situaciones es adecuado aplicar los estimadores de la selección purificadora que se han desarrollado en esta tesis, hemos llevado a cabo simulaciones. Se han contemplado dos escenarios: (1) simulaciones donde asumimos que los sitios segregan independientemente entre sí (libre recombinación) y el censo efectivo es constante (véase sección 2.3.1). Estas simulaciones han servido para estudiar el efecto de distintas *DFEs* (*distribution of fitness effects*) y número de sitios segregantes sobre los nuevos estimadores (véase sección 3.1.2). (2) Simulaciones donde los sitios no segregan libremente (están ligados) y el censo efectivo fluctúa a lo largo del tiempo (véase sección 2.3.2). Estas últimas simulaciones han ayudado a conocer el efecto combinado de la interferencia entre sitios y la demografía sobre nuestros estimadores (véase sección 3.1.2).

### 2.3.1 LIBRE RECOMBINACIÓN ENTRE SITIOS Y CENSO EFECTIVO CONSTANTE

Para aplicar nuestros estimadores necesitamos simular el SFS de sitios selectivos y neutros para un tamaño de muestra, longitud de secuencia y tasa de mutación dada. Todas las simulaciones de esta sección han sido computadas mediante ecuaciones de difusión, mientras que las simulaciones de la siguiente sección han sido computadas mediante un simulador *forward in time*.

Para obtener el SFS simulado necesitamos computar primero el SFS esperado. El término esperado se refiere a lo esperado en el caso donde el número de sitios segregantes del que disponemos para realizar las estimas es infinito. Una vez tenemos el SFS esperado, el SFS simulado se obtiene muestreando un número limitado de sitios segregantes al azar (véase detalles más abajo). El SFS esperado en la clase de sitios seleccionados ha sido computado utilizando la aproximación de difusión de Kimura (1983) donde el efecto de las nuevas mutaciones deletéreas sobre la *fitness* (esto es la *DFE*) se ha modelado a partir de una distribución gamma (la parte negativa de esta). Estas simulaciones están basadas en el trabajo de Eyre-Walker *et al.* (2006) donde  $P_n(j)$  es el número de mutaciones (o SNPs) seleccionados a frecuencia  $j$  en una muestra de tamaño  $n$ :

$$P_n(j) = 2N_e u L_n \int_{-\infty}^{\infty} \int_0^1 D(\varphi, \beta, s) H(N_e, s, x) Q(n, j, x) dx \cdot ds \quad (2.1)$$

donde  $u$  es la tasa de mutación por sitio,  $N_e$  es el tamaño de población efectivo,  $L_n$  es el número de sitios seleccionados y

$$D(\varphi, \beta, s) = \frac{\varphi^\beta s^{\beta-1} e^{-\varphi s}}{\Gamma(\beta)}, \quad (2.2)$$

es la *DFE* de las nuevas mutaciones, la cual asumimos es una distribución gamma con parámetro de forma, o *shape parameter*,  $\beta$ .  $\varphi$  es el *location parameter* que guarda relación con la media de la distribución o el efecto promedio de las nuevas mutaciones sobre la eficacia biológica,  $\gamma = \overline{N_e s} = \beta/\varphi$ .

$$H(N_e, s, x) = 2 \left( \frac{1-e^{4N_e s(1-x)}}{x(1-x)(1-e^{4N_e s})} \right) \quad (2.3)$$

$H(N_e, s, x)$  es el tiempo que una nueva mutación semidominante con coeficiente de selección  $s$  (en el heterocigoto) pasa entre la frecuencia  $x$  y la frecuencia  $x + dx$  (Wright 1938a). Finalmente,

$$Q(n, j, x) = \begin{cases} \frac{n!}{j!(m-j)!} (x^j(1-x)^{n-j} + x^{n-j}(1-x)^j) & \text{si } j \neq n/2 \\ \frac{n!}{j!(m-j)!} (1-x)^{n-j} & \text{si } j = n/2 \end{cases} \quad (2.4)$$

es la probabilidad de observar una mutación a frecuencia  $x$  en  $j$  de  $n$  secuencias. Suponemos que las nuevas mutaciones beneficiosas contribuyen cuantitativamente muy poco al SFS de los sitios seleccionados, pues permanecen poco tiempo segregando en las poblaciones (véase figura 1.2), por esta razón no se han modelado sus efectos.

El SFS esperado de la clase de sitios neutros es mucho más sencillo de computar.  $P_s(j)$  es el número de SNPs neutros a frecuencia  $j$  en una muestra de tamaño  $n$  (Eyre-Walker *et al.* 2006):

$$P_s(j) = 4N_e u L_s \left( \frac{1}{j} + \frac{1}{n-j} \right) \quad (2.5)$$

$L_s$  es el número de sitios neutros en la secuencia. Tanto para el SFS de la clase de sitios neutros como de la de sitios seleccionados desconocemos el estado ancestral y derivado de las mutaciones, por lo tanto, un alelo de un SNP dialélico segregando a frecuencia  $x$  es equivalente a un alelo de un SNP segregando a frecuencia  $1 - x$ . A este tipo de SFS se lo representa como MAF (*minor allele frequency*) o *folded-SFS*.

Para generar los SFS simulados necesitamos muestrear al azar un número limitado de sitios segregantes a partir del SFS esperado. Para hacer esto primero normalizamos el SFS neutro (y selectivo) esperado de modo que el sumatorio de todos los alelos sea igual a 1. Es importante destacar que el tamaño de muestra ( $n$ ) utilizado para computar el SFS esperado y simulado es de 128 cromosomas o individuos haploides. Esto es así porque este ha sido el número de cromosomas que hemos utilizado para estimar el SFS en la población natural de *D. melanogaster* en la que se ha centrado esta tesis (véase sección 2.2.3). Segundo, generamos una variable

aleatoria  $X$  a partir de una distribución de Poisson con media  $\theta L_s a$ . En nuestro caso, por ejemplo, un gen promedio contiene 14 sitios segregantes sinónimos 4-veces degenerados (estimas basadas en una población norteamericana de *D. melanogaster* donde 128 cromosomas fueron muestreados). Por lo tanto,  $\theta L_s a = 14$ , donde  $L_s$  es el número de sitios sinónimos,  $\theta$  es la tasa de mutación por sitio escalada por el tamaño poblacional efectivo ( $\theta = 4N_e u$ ) y  $a$  es la corrección de Watterson ( $a$  es la suma de  $1/i$  desde  $i = 1$  a  $i = n-1$ ) (Charlesworth y Charlesworth 2010; p. 29). Finalmente, para cada valor de la variable  $X$  muestrearemos el número de sitios segregantes correspondiente en proporción a la frecuencia de dichos alelos en el SFS normalizado. Es decir, si quisiéramos muestrear un único sitio segregante, este es más probable que esté a baja frecuencia que a frecuencia intermedia, simplemente porque en el SFS normalizado hay más alelos a baja frecuencia que a frecuencia intermedia. A su vez, para generar el SFS simulado de sitios selectivos hay que tener en cuenta que la selección purificadora disminuye el número de sitios segregantes selectivos por sitio selectivo ( $p_N$ ) en relación al número de sitios segregantes neutros por sitio neutro ( $p_S$ ) y además que la relación sitio sinónimo - sitio no-sinónimo es de 1 a 2 de acuerdo al código genético ( $L_n/L_s = 2$ ). Por lo tanto, el número de sitios segregantes promedio en el caso de las mutaciones seleccionadas corresponde a  $2p_N/p_S \times \theta L_s a$ , donde  $p_N$  y  $p_S$  provienen del sumatorio de los alelos de los respectivos SFS de las clases de sitios seleccionados y neutros esperados.

Se ha estudiado el comportamiento de nuestros estadísticos (definidos en la sección de resultados 3.1) cuando el número de sitios segregantes del que disponemos es limitado; esto sucede comúnmente cuando queremos realizar estimas a nivel de un solo gen. No es lo mismo disponer de 5 sitios segregantes que de 500, pues las estimas basadas en un número mayor de sitios segregantes estarán sujetas a menor error estadístico y serán más fiables. Por lo tanto, nos hemos preguntado con cuantos sitios segregantes es seguro aplicar nuestros estimadores. Para responder a esta

pregunta hemos calculado por un lado el valor promedio de los estadísticos utilizando un número limitado de sitios segregantes y por otro lado hemos calculado la desviación estándar de la media:

$$s_{\bar{x}} = \sqrt{\frac{1}{n} \sum_{i=1}^n (E - O)^2} \quad (2.6)$$

Donde  $E$  es el valor esperado (o verdadero) de los estadísticos, el cual definimos como el valor obtenido utilizando infinitos sitios segregantes,  $O$  es el valor observado con un número limitado de sitios segregantes (donde se ha aplicado la corrección +1 en el denominador [véase sección 3.1]) y  $n$  en este caso es el número de iteraciones sobre una misma combinación de parámetros. Para este estudio  $n = 100$ . Hemos probado 3 valores promedio de sitios segregantes neutros: (1) los encontrados a nivel de un gen prototípico de *D. melanogaster*, es decir 14 sitios segregantes 4-veces degenerados, (2) los encontrados concatenando 10 de estos genes (o 140 sitios segregantes) y (3) los encontrados concatenando 100 de estos genes (o 1400 sitios segregantes). El número y el SFS de la clase de alelos seleccionados dependen de la *DFE*, por lo tanto, no hemos impuesto ningún límite al respecto.

En resumen, estas simulaciones nos han servido para conocer el efecto de distintas *DFEs* sobre nuestros estimadores teniendo en cuenta consideraciones prácticas como el número de sitios segregantes disponibles.

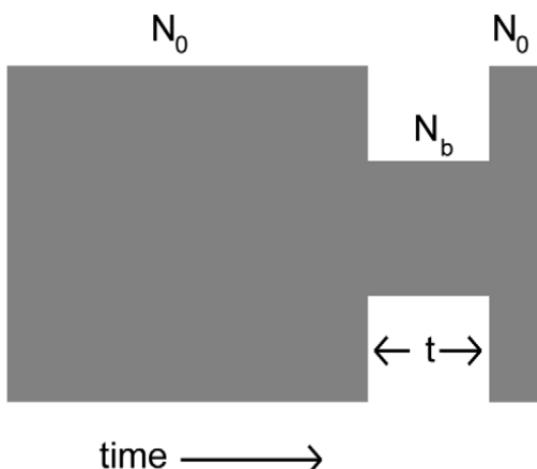
### 2.3.2 SITIOS LIGADOS Y CENSO EFECTIVO VARIABLE

El objetivo de estas simulaciones es estudiar cómo afecta a nuestros estadísticos la variación en la tasa de recombinación entre alelos y un cuello de botella reciente asumiendo la *DFE*, tasa de mutación y densidad génica estimada en *D. melanogaster*.

Hemos utilizado simulaciones *forward in time* porque actualmente no hay forma analítica de simular cambios en el tamaño de población en presencia de selección (y recombinación). El *software* utilizado para computar el SFS de sitios selectivos y neutros y con ello investigar el comportamiento de nuestros estimadores bajo un cuello de botella severo reciente (como el experimentado por nuestra población norteamericana de *D. melanogaster*) ante distintas tasas de recombinación se denomina SFS\_CODE (*Selection on Finite Sites under Complex Demographic Events*, Hernandez 2008). Este programa es capaz de ejecutar simulaciones *forward* bajo una gran variedad de modelos demográficos y distribuciones de coeficientes de selección, entre muchas otras opciones. Este programa de simulación asume el modelo Wright-Fisher (véase sección 1.1.1), siguiendo un modelo de sitios finitos con selección, recombinación (entre cruzamiento y/o conversión génica) y demografía para un número arbitrario de poblaciones. La población entera se sigue generación tras generación desde el inicio de la simulación hasta el momento del muestreo a diferencia de lo que ocurre en simulaciones de coalescencia, donde la historia de una muestra se simula hacia atrás en el tiempo hasta su fundador (véase cuadro 2). El programa genera por tanto una muestra de DNA, un alineamiento, en el que un sitio ha podido mutar, seleccionarse y/o recombinar con otros varias veces a lo largo de la simulación. En este caso no tenemos que hacer el paso del SFS esperado al SFS simulado, pues obtenemos directamente el SFS simulado a partir del alineamiento.

Para este trabajo hemos simulado 100 *loci* codificadores de 1.25 kb separados por 9 kb de secuencia (neutra) no-codificadora, esta densidad génica equivaldría a la encontrada en una región de 1,025 Mb del genoma de *D. melanogaster* (Adams *et al.* 2000). La variación en la tasa de recombinación (entre cruzamiento en este caso) ha sido simulada variando el valor del parámetro  $\rho$  ( $\rho = 4N_e r$ ). Seis valores distintos de  $\rho$  han sido estudiados: 0, 0'001, 0'01, 0'1, 0'2, 0'4. La diversidad nucleotídica neutra esperada  $\Theta$  ( $\Theta = 4N_e u$ ) es de 0,01 y equivale al valor promedio estimado en *D. melanogaster* en norteamericana (Singh *et al.* 2007; Sackton *et al.* 2009; Mackay *et al.*

2012; Langley *et al.* 2012; Pool *et al.* 2012). El tamaño de población simulado es de 1000 individuos diploides con un *burn-in* inicial de 10.000 generaciones, este *burn-in* es muy importante, permite que la diversidad genética de las simulaciones no parte de 0 sino del valor esperado bajo el equilibrio mutación-selección-deriva. El tamaño de muestra ( $n$ ) utilizado para extraer el SFS simulado es de 128 cromosomas o individuos haploides. Los valores de los parámetros de la *DFE* son  $\beta = 0,3$  ( $\beta$  es el parámetro de forma de la distribución gamma asumida para modelar la *DFE*) y  $\gamma = \overline{N_e s} = 1000$  ( $\gamma$  es el efecto promedio de las nuevas mutaciones deletéreas). Estos son los parámetros de la *DFE* para nuevas mutaciones no-sinónimas estimados por Keightley y Eyre-Walker (2007) en *D. melanogaster*. Asumimos que ninguna mutación (nueva o preexistente) es adaptativa. En las simulaciones con cuellos de botella los parámetros estimados por Thornton y Andolfatto (2006) fueron los utilizados. La figura 2.3 muestra el valor de dichos parámetros.



**FIGURA 2.3** Modelo demográfico de referencia. Siguiendo la Tabla 2 del trabajo de Thornton y Andolfatto (2006) para una razón entre la tasa poblacional de recombinación y mutación,  $\rho/\theta=10$  (esta es la relación promedio genómica), la población alcanza el equilibrio en un tamaño poblacional  $N_0$ , se contrae a un tamaño  $N_b$ , y luego se vuelve a expandir al tamaño  $N_0$  después de  $4N_0t$  generaciones. La población permanece  $4N_0$  [0,0042] generaciones antes del muestreo actual. En nuestro modelo  $N_0 = 1000$  individuos diploides,  $N_b = 29$  y  $4N_0t = 0,015$ . [Figura tomada de Feder *et al.* (2012)].

El SFS neutro (y selectivo) ha sido estimado concatenando directamente todos los sitios segregantes sinónimos (y no-sinónimos) de los 100 *loci* codificadores. Se han realizado 100 iteraciones para cada una de las 6 tasas de recombinación estudiadas y sobre estas 100 iteraciones se ha calculado el valor promedio y desviación estándar de los estadísticos con tal de conocer como la interferencia entre sitios y la demografía afectan a nuestras estimas en relación a lo esperado cuando los sitios son independientes y el tamaño efectivo de población permanece constante. En este caso no es posible conocer el valor esperado con infinitos sitios, esto implicaría simular un cromosoma de tamaño infinito (computacionalmente imposible).

Ejemplo de instrucción para una simulación en ausencia de recombinación y un cuello de botella:

```
./sfs_code 1 100 -o fly --seed 10092015 --popSize 1000 -n 128 -t 0.01 -r 0.0 -W 2 0.00 1 1 0.3 0.0003  
-L 200 1250 9000 R -l g 0 R -a C N R -Td 0 0.029 -Td 0.015 34.48 -TE 0.0192 -noSeq
```

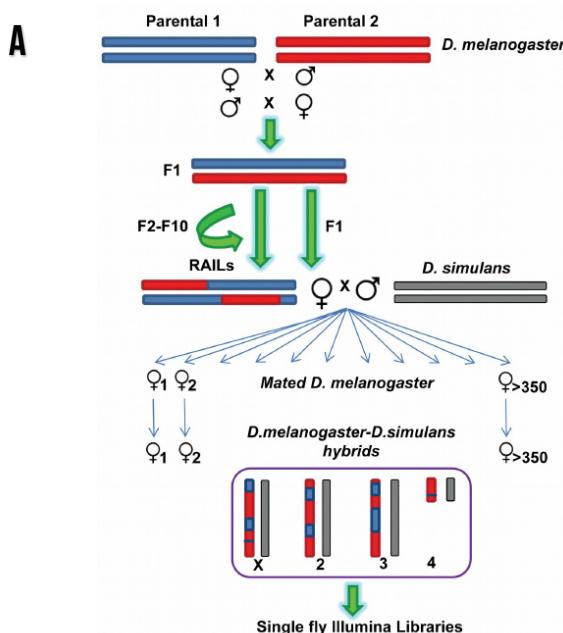
Para extraer el SFS de sitios selectivos (y neutros) del *output* anterior (denominado *fly*) para una muestra de  $n = 128$  utilizamos el programa convertSFS\_CODE (Hernandez 2008).

## 2.4 ESTIMACIÓN DE LA TASA DE RECOMBINACIÓN

La tasa de recombinación a escala fina ( $C_{FS}$ , del inglés *Fine-Scale*) utilizada en este trabajo es una estima de alta resolución, es la primera descripción empírica de la recombinación genómica a escala molecular obtenida en una especie (Comeron *et al.* 2012). Los autores afirman que el mapa es 50 veces más detallado que el mapa de entrecruzamiento de alta resolución actual del genoma humano. En este trabajo hemos utilizado la tasa de entrecruzamientos (no la conversión génica) en ventanas no solapantes de 100 kb las cuales están disponibles públicamente en la dirección URL <http://bioweb.biology.uiowa.edu/labs/comeron/recombination>. Los marcadores genéticos utilizados son SNPs y pequeños INDELS obtenidos de un subconjunto de 10 cepas de Carolina del Norte además de dos cepas africanas. El límite del tamaño de la ventana en nuestro análisis viene determinado por estas estimas de recombinación que sólo están disponibles en ventanas de 100 kb. Para más detalles sobre la obtención de las estimas de recombinación y un resumen de los resultados más relevantes del trabajo de Comeron *et al.* (2012) consultese cuadro 5.

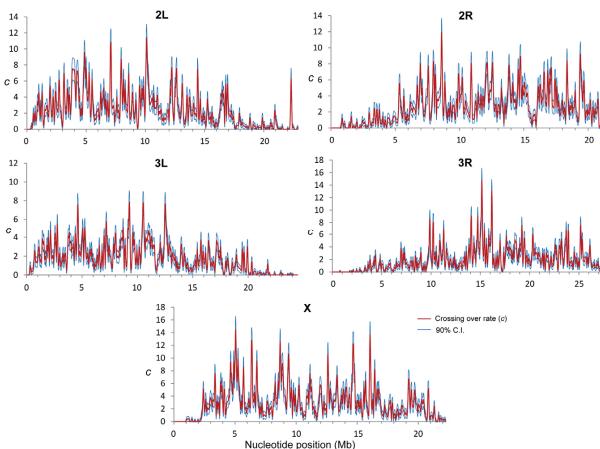
Para los análisis basados en ventanas cromosómicas no solapantes de 1 Mb, la tasa de recombinación de la ventana corresponde al promedio de las 10 ventanas de 100 kb que la integran. En cambio, para los resultados basados en agrupaciones de genes, cada gen de nuestro conjunto de datos tiene asignada la tasa de recombinación (en cM/Mb) de la ventana cromosómica en la que se encuentra localizado – la coordenada de los genes es única y corresponde a la posición intermedia entre el primer nucleótido del primer exón y último nucleótido del último exón de dicho gen. La recombinación asignada al grupo de genes es el promedio de la recombinación asignada a cada uno de los genes.

## CUADRO 5: CARTOGRAFÍA GENÓMICA DE LA RECOMBINACIÓN A ALTA RESOLUCIÓN EN *D. melanogaster*



Comeron *et al.* (2012) estiman la tasa de recombinación a alta resolución en *D. melanogaster* a nivel genómico en distintos individuos y distinguiendo entre entrecruzamiento y conversión génica. Para ello, se llevaron a cabo 8 cruces recíprocos entre 10 líneas del proyecto DGRP, una línea de Madagascar y otra de Papúa Nueva Guinea. De los descendientes de estos cruces (F1) se realizaron cruces hermano-hermana para generar las líneas homocigóticas denominadas RAIL (*Recombinant Advanced Intercross Lines*) (ver panel A). Las hembras RAIL se cruzaron con machos *D. simulans*. Finalmente, las hembras híbridas de *D. melanogaster* se secuenciaron, se eliminaron las lecturas potenciales de *D. simulans* y se seleccionaron aquellas que cartografiaban únicamente en uno de las dos líneas parentales. Así, se pudieron obtener los haplotipos (potencialmente recombinantes) de cada brazo cromosómico de las distintas hembras híbridas de cada cruce (ver panel A) [esquema de Comeron *et al.* (2012)].

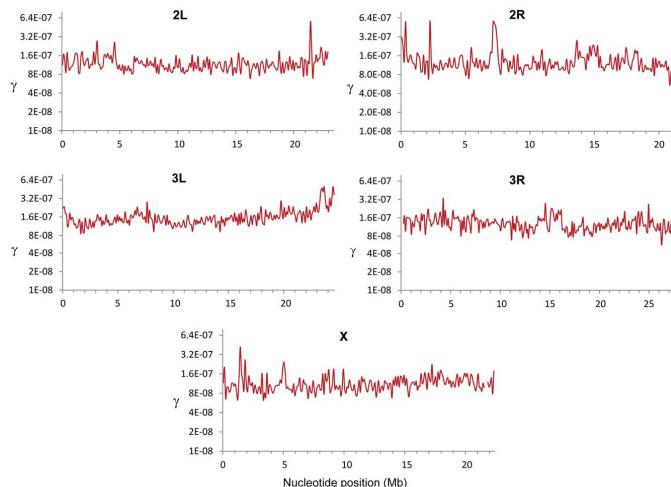
**B**



La tasa de entrecruzamiento muestra el patrón general previamente conocido con una reducción paulatina desde el inicio del tercio correspondiente a la región centromérica avanzando en la dirección hacia el centrómero y una drástica reducción en la proximidad del telómero (ver panel B, donde  $c = \text{cM/Mb}$  en ventanas no solapantes de 100 kb). Sin embargo, se observan que en las regiones previamente consideradas de alta recombinación (distintas de las centroméricas o teloméricas) las tasas de recombinación varían entre 15 y 20 veces entre ellas. Es decir, que hay una gran heterogeneidad intracromosómica para la tasa de entrecruzamiento. Además, la tasa de entrecruzamiento también muestra heterogeneidad entre individuos o cruces (véase figura 2 de Comeron *et al.* [2012]).

## CUADRO 5: CONTINUACIÓN

C



El estudio de la conversión génica mostró por otro lado que la distribución de la tasa de conversión génica a lo largo de los cromosomas es mucho más homogénea que la tasa de entrecruzamiento, sin aparente disminución en las regiones centroméricas y teloméricas (ver panel C, donde  $\gamma = /pb/meiosis$  hembra en ventanas no solapantes de 100 kb). Además, los autores indican que la longitud de los fragmentos de conversión génica son en promedio de 518 pb y el ~83% de los sucesos iniciales de recombinación homóloga se resuelven como conversiones génicas, es decir, la conversión génica es el tipo de recombinación mayoritaria en *D. melanogaster*.

## 2.5 ESTIMACIÓN DE LA DENSIDAD GÉNICA

Para estimar la densidad génica de la ventana donde yace un gen hemos utilizado todos los sitios codificadores presentes en nuestro archivo de anotaciones (versión 5.50). Primero hemos calculado la coordenada central de los genes de nuestro conjunto de datos, la cual corresponde a la posición intermedia entre el primer nucleótido del primer exón y último nucleótido del último exón de dicho gen. Finalmente, hemos contado todos los sitios codificadores 50 kb aguas arriba y 50 kb aguas abajo de la coordenada central. Cada gen tiene su propia estima de densidad génica en ventanas de 100 kb. Esta densidad génica se ha correlacionado con la tasa de adaptación (véanse secciones 3.3.2 y 3.3.4) con tal de averiguar si la iHR es mayor en aquellos genes sumergidos en regiones de alta densidad génica.

## 2.6 ESTIMACIÓN DE LA TASA DE MUTACIÓN

Para analizar el papel de la tasa de mutación sobre la tasa de adaptación (véase sección de resultados 3.3.3), hemos correlacionado la divergencia en las posiciones 4-veces degeneradas ( $K_4$ ) con la tasa de adaptación no-sinónima ( $K_{a+}$ ) (definido en la siguiente sección). Sin embargo,  $K_4$  y  $K_{a+}$  no son variables independientes, la estima de  $K_{a+}$  depende de  $K_4$  (consúltese sección 2.7 expresión [2.14]), por lo tanto, esperamos que  $K_4$  y  $K_{a+}$  estén negativamente correlacionadas simplemente debido al error de muestreo en  $K_4$ . Para deshacernos de esta falta de independencia hemos obtenido tres estimas independientes de  $K_4$  a partir de tres muestreos de las posiciones 4-veces degeneradas generado mediante una variable aleatoria hipergeométrica multivariante (este sencillo “truco” estadístico es similar al aplicado anteriormente en los trabajos de Smith y Eyre-Walker 2002; Piganeau y Eyre-Walker 2009; Stoletzki y Eyre-Walker 2011; Gossman *et al.* 2012).

La distribución hipergeométrica sirve para modelizar procesos de *Bernouilli* con probabilidades no constantes (sin reemplazamiento). La distribución hipergeométrica es especialmente útil en todos aquellos casos en los que se extraigan muestras de poblaciones finitas o se realizan experiencias repetidas sin devolución del elemento extraído o sin retornar a la situación experimental inicial. Modeliza, de hecho, situaciones en las que se repite un número determinado de veces una prueba dicotómica de manera que con cada sucesivo resultado se ve alterada la probabilidad de obtener en la siguiente prueba uno u otro resultado. Es una distribución fundamental en el estudio de muestras pequeñas de poblaciones (también) pequeñas y en el cálculo de probabilidades de juegos de azar. En nuestro caso, imaginemos a los genes como urnas donde el número de posiciones 4-veces degeneradas que han sufrido una fijación (o substitución) se encuentran representadas como bolas negras (que denominamos  $D_4$ ), mientras que el resto de posiciones 4-veces degeneradas que no han sufrido ningún cambio desde la separación de ambas especies se encuentra representado por bolas blancas (que denominamos  $L_4 - D_4$ ). La suma

de bolas negras y blancas son el total de posiciones 4-veces degeneradas en dicho gen (o  $L_4$ ). Para cada gen hemos muestreado  $n$  posiciones 4-veces degeneradas al azar sin reemplazamiento tres veces consecutivas utilizando la función de probabilidad de una variable aleatoria con distribución hipergeométrica:

$$P(X = x) = \frac{\binom{D_4}{x} \binom{L_4 - D_4}{n-x}}{\binom{L_4}{n}} \quad (2.7)$$

donde  $n$  es el tamaño de la muestra extraída (en este caso  $n = L_4/3$ ),  $D_4$  es el número de elementos en la población original que pertenecen a la categoría deseada y  $x$  es el número de elementos en la muestra que pertenecen a dicha categoría. En concreto hemos generado las tres estimas de  $K_4$  de la siguiente manera:

$$D_{4,1} = \text{multivariateHypergeometric}(D_4, 0.33 \times L_4), \quad (2.8)$$

$$D_{4,2-3} = D_4 - D_{4,1}, \quad (2.9)$$

$$D_{4,2} = \text{multivariateHypergeometric}(D_{4,2-3}, 0.33 \times L_4), \quad (2.10)$$

$$D_{4,3} = D_4 - D_{4,1} - D_{4,2} \quad (2.11)$$

Donde  $D_{4,1}$ ,  $D_{4,2}$  y  $D_{4,3}$  son el número de sitios 4-veces degenerados divergentes después de elegir al azar un tercio de todas las posiciones 4-veces degeneradas del gen tres veces consecutivas y sin reemplazamiento. Hemos dividido  $D_{4,1}$ ,  $D_{4,2}$  y  $D_{4,3}$  por  $1/3 \times L_4$ , para obtener  $K_{4,1}$ ,  $K_{4,2}$  y  $K_{4,3}$ , respectivamente.  $K_{4,1}$  ha servido para ordenar los genes y asignarlos a grupos de acuerdo a su tasa de mutación,  $K_{4,2}$  se ha utilizado para estimar la tasa de substituciones no-sinónimas adaptativas ( $K_{\theta+}$ ) y  $K_{4,3}$  ha sido nuestra estima de la tasa de mutación, la cual hemos correlacionado con  $K_{\theta+}$ .

Queremos probar si los genes con elevadas tasas de mutación han perdido más substituciones adaptativas debido a la interferencia de Hill-Robertson (iHR). Para hacer esto hemos dividido nuestro conjunto de datos en dos mitades iguales de acuerdo a su tasa de mutación. De nuevo, con tal de evitar que el error de muestreo en nuestras estimas de  $K_4$  genere una correlación negativa con nuestras estimas de

$K_{\alpha+}$ , y con ello una subestima de la fracción de substituciones adaptativas perdidas debido a la iHR, hemos estimado dos variables independientes de  $K_4$ :  $K_{4,1}$  se ha usado para ordenar y categorizar los genes de acuerdo a su tasa de mutación y  $K_{4,2}$  se ha utilizado para estimar  $K_{\alpha+}$ . En este caso el número de posiciones 4-veces degeneradas muestreadas utilizando la distribución hipergeométrica para cada gen es  $n = L_4/2$ :

$$D_{4,1} = \text{multivariateHypergeometric}(D_4, 0.5 \times L_4), \quad (2.12)$$

$$D_{4,2} = D_4 - D_{4,1} \quad (2.13)$$

Hemos dividido  $D_{4,1}$  y  $D_{4,2}$  por  $1/2 \times L_4$  para obtener  $K_{4,1}$  y  $K_{4,2}$ , respectivamente.

## 2.7 ESTIMACIÓN DE LA DFE Y DE LA TASA DE EVOLUCIÓN ADAPTATIVA

Para estimar la *DFE* de nuevas mutaciones deletéreas (Keightley y Eyre-Walker 2007) y la tasa de evolución adaptativa (Eyre-Walker y Keightley 2009) hemos utilizado el paquete de programas contenidos en la versión descargable de *DFE-alpha* (versión 2.14, disponible en URL: <http://www.homepages.ed.ac.uk/pkeightl/software>). Las estimas basadas en genes individuales son a menudo indefinidas por la falta de sitios segregantes (o divergentes), de modo que ha sido necesario combinar los datos de múltiples *loci* para estimar la *DFE* y la tasa de evolución adaptativa. Por un lado, hemos realizado estimas a nivel de ventanas de 1 Mb (las cuales disponen de información suficiente como para evitar estimas indefinidas), y por otro hemos agrupado los genes en grupos de más de 100 genes codificadores para evitar este problema. Hemos ejecutado el programa para cada ventana o grupo de genes de manera independiente.

Para aplicar este método es imprescindible tener una secuencia de referencia neutra que sirva de hipótesis nula para determinar la tasa de substituciones adaptativas y la *DFE* en una secuencia diana particular (en este estudio esta secuencia diana son

los sitios UTR, intergénicos, intrónicos y codificadores 0-veces degenerados). Las mutaciones puntuales que ocurren en las posiciones 4-veces degeneradas se han considerado el mejor indicador (*proxy*) de la tasa y el espectro de frecuencias que esperaríamos encontrar en mutaciones puramente o efectivamente neutras. Aunque para algunos análisis complementarios (véase sección 3.3.1) los intrones cortos de menos de 66 pb también han sido utilizados como referencia de neutralidad.

En la sección 1.3.3 se explicó en detalle el fundamento de los métodos de tipo *DFE-alpha* que combinan la inferencia de la demografía y la *DFE* estimada mediante máxima verosimilitud con las estimas de la tasa de evolución adaptativa. Brevemente, se dijo que *DFE-alpha* modela la *DFE* de nuevas mutaciones deletéreas que ocurren en sitios putativamente funcionales asumiendo una distribución gamma con dos parámetros: (1) el efecto promedio de las nuevas mutaciones sobre la eficacia biológica,  $\gamma = -N_e s$ , y (2) un parámetro de forma  $\beta$  que permite que la distribución pueda adquirir una gran variedad de formas, desde la distribución uniforme hasta la exponencial. *DFE-alpha* puede modelar también un cambio instantáneo (en una sola generación), que ocurrió hace  $t_2$  generaciones, del tamaño poblacional ancestral  $N_1$  al tamaño poblacional actual  $N_2$  utilizando la información almacenada en el espectro de frecuencias de mutaciones putativamente neutras. *DFE-alpha*, por tanto, infiere  $\gamma$ ,  $\beta$ ,  $N_2/N_1$  y  $t_2$ , para dar estimas prácticamente insesgadas de la fracción de substituciones adaptativas ( $\alpha$ ) en los sitios funcionales. Esto se consigue integrando la información proveniente de la divergencia de sitios putativamente funcionales (por ejemplo, los sitios 0-veces degenerados,  $K_a$ ) y neutros (como los sitios 4-veces degenerados,  $K_s$ ), y sus respectivos espectros de frecuencias, los cuales deben ser previamente calculados por el usuario (véase sección 2.3.2). Por lo tanto, la estima de  $\alpha$  viene definida por la siguiente ecuación:

$$\alpha = \frac{K_a - K_4 \int_0^{\infty} 2Nu(N,s)f(s|\gamma,\beta)ds}{K_a} \quad (2.14)$$

Donde  $u(N, s)$  es la probabilidad de fijación de una nueva mutación funcional con coeficiente de selección  $Ns$  (en este caso  $N = N_e$ ) que aparece en copia única en una población diploide de tamaño efectivo  $N$  (Kimura 1957, 1983) (consúltese expresión [1.2] en sección 1.1.1).  $f(s|\gamma, \beta)$  es la función gamma utilizada para modelar la *DFE* (consúltese expresión [2.2]). Nuestras estimas de  $\alpha$  dependen tanto de la probabilidad de fijación de las nuevas mutaciones efectivamente neutras y deletéreas, esto es, la *DFE* y el efecto que tiene sobre ella la demografía, y la tasa de mutación neutra (en nuestro caso utilizamos  $K_4$  como indicador). A partir de la estima de  $\alpha$  es sencillo obtener otros dos estadísticos comúnmente utilizados para medir la tasa de evolución adaptativa, estos son: (1) el número de substituciones adaptativas por sitio funcional ( $K_{\alpha+} = \alpha \times K_\alpha$ ) y (2) el número de substituciones adaptativas en relación al número de substituciones neutras, o la tasa de mutación ( $\omega_A = K_{\alpha+} / K_4$ ) (Gossman *et al.* 2010).

A parte de estimar la tasa de adaptación (utilizando  $\omega_A$  en este caso) también se dan estimas de la proporción de mutaciones deletéreas con coeficientes de selección en distintos rangos de  $N_e s$ , estos son:  $-1 < N_e s < 1$  (efectivamente neutras),  $-10 < N_e s < -1$  (ligeramente deletéreas) y  $N_e s < -10$  (fuertemente deletéreas). Esto se ha realizado a través del programa *prop\_muts\_in\_s\_ranges* el cual forma parte del paquete de programas de *DFE-alpha*. Para estimar la *DFE* de las nuevas mutaciones deletéreas se ha utilizado el programa *est\_DFE* y para estimar la tasa de adaptación se ha utilizado el programa *est\_alpha\_omega* ambos programas se encuentran integrados en el paquete *DFE-alpha*.

## 2.8 ESTIMACIÓN DEL SESGO EN EL USO DE CODONES

El software *CodonW* (<http://codonw.sourceforge.net/>) (Peden 1999) ha sido utilizado para estimar un índice que mide la frecuencia de los codones óptimos dentro de un gen, este es *Fop* (*frequency of optimal codons*) (Ikemura 1981; Ikemura 1985; Ikemura y Ozeki 1982). Es una medida especie específica que mide el sesgo hacia aquellos codones que parecen ser óptimos desde el punto de vista de la traducción. Es una sencilla razón entre el número de codones óptimos y el total de codones. Su valor va de 0 (cuando un gen no contiene codones óptimos) a 1 (cuando todos los codones de un gen son óptimos). Para calcular *Fop* es necesario conocer previamente el conjunto de codones óptimos para cada aminoácido, esto es lo que hicieron Shields *et al.* (1988). *CodonW* dispone de esta información por defecto así que para estimar *Fop* sólo se requiere disponer de la secuencia de ADN en formato fasta para cada gen. Estas secuencias fueron descargas de <http://flybase.org/> (versión 5.50).

## 2.9 GENES DEL SISTEMA INMUNE Y EXPRESIÓN SESGADA EN MACHOS

Si los genes que se han descrito que tienen elevadas tasas de adaptación en *Drosophila* como los involucrados en el sistema inmune y/o de expresión sesgada en machos (implicados en la producción de esperma o específicos de testículos por ejemplo) (Obbard *et al.* 2009; Pröschel *et al.* 2006; Haerty *et al.* 2007), se encuentran sobrerepresentados en aquellas agrupaciones de genes con mayor (o menor) tasa de recombinación (véase sección 3.3.1), densidad génica (véase sección 3.3.2) o tasa de mutación (véase sección 3.3.3), entonces las correlaciones que encontramos entre estas variables y la tasa de adaptación podrían no depender ni de la recombinación, ni de la densidad génica, ni de la tasa de mutación diferencial entre grupos. Luego, descargamos los términos *Gene Ontology* (GO) para el conjunto de 6.141 genes de la sección de resultados 3.3 desde el *release 78* de *Fruitfly* utilizando el paquete de *R* *biomaRT* (Durinck *et al.* 2005). Creamos una lista de términos GO relacionados con la respuesta inmune, testículos y producción de esperma utilizando la herramienta

*QuickGO* (Binns *et al.* 2009). Cuando un gen determinado tenía un término GO presente en la lista anterior se etiquetaba como gen de sistema inmune y/o expresión sesgada en machos, el resto de genes se etiquetaron como genes control. La lista de términos GO que se utilizó para etiquetar a los genes se puede encontrar en la tabla S6 del ANEXO.

## 2.10 ANÁLISIS ESTADÍSTICO

Todos los análisis estadísticos se realizaron con el paquete estadístico R (versión 3.0.2) (*R Core Team* 2013) y en algunos casos la preparación de los datos de entrada se pre-procesaron con *Bash* y *Perl scripting* (véase sección 2.11). R es un lenguaje y entorno para computación y gráficos estadísticos. Es un proyecto GNU, es similar al lenguaje S y se ha desarrollado en los Laboratorios Bell por John Chambers y colegas (<https://www.r-project.org/>).

### 2.10.1 CORRELACIONES, REGRESIONES Y ANCOVA

Se ha calculado el coeficiente para la correlación de *Spearman* ( $\rho_s$ ) utilizando la función de R “*cor.test*” (del paquete de R *base*). La correlación de *Spearman* mide estadísticamente la interdependencia entre dos variables aleatorias continuas utilizando una función monotónica. Para calcular  $\rho_s$ , los datos son ordenados y reemplazados por su respectivo orden. Es decir, en esta prueba no importa tanto el valor de los datos como su orden. Es una alternativa no paramétrica a la correlación de *Pearson*, pero más robusta a desviaciones de la linealidad porque no asume una relación lineal entre las dos variables. Como alternativa no paramétrica a la prueba t de *Student* y con tal de probar si, bajo la hipótesis nula, la distribución de partida de dos grupos (o muestras) independientes es la misma se utilizó la prueba de Mann-Whitney U. En esta prueba, bajo la hipótesis alternativa, los valores de una de las dos muestras tienden a exceder a los de la otra. El valor p de la prueba se calculó utilizando la función de R “*wilcox.test*” (del paquete de R *base*). En este estudio se

han realizado tres tipos de regresiones: lineales, no-lineales y locales, a lo largo de los distintos apartados. Las dos primeras fueron ejecutadas utilizando la función de R “*nls*” (del paquete de R *stats*) mientras que para la regresión local se utilizó la función “*loess*” (del paquete de R *stats*). Para la regresión local aumentamos el parámetro *span* de 0,75 a 1. Aumentar el parámetro *span* disminuye la suavidad (o *smoothness*) de la curva y aumenta, por tanto, la robustez entre réplicas del *bootstrap*. Para comprobar si la función lineal explica mejor (o no) la relación entre la tasa de recombinación y  $K_{a+}$  que la función no lineal utilizamos la función de R “*ANOVA*” (del paquete de R *base*) el cual computa el *Akaike Information Criterion* (AIC) para cada modelo y el valor p del modelo con más parámetros respecto al modelo más sencillo. Tanto el análisis de la covarianza (ANCOVA) como las regresiones múltiples se han llevado a cabo invocando la misma función de R, “*lm*” (del paquete de R *base*). Para generar la variable aleatoria hipergeométrica de la sección 2.6 hemos utilizado la función de R “*rhyper*” (del paquete de R *stats*). En las figuras de este trabajo los valores p entre 0,05 – 0,01 se representan con un asterisco, los valores p entre 0,01 – 0,001 se representan con dos asteriscos y los valores p menores a 0,001 se representan con tres asteriscos.

## 2.10.2 ESTIMACIÓN DEL EFECTO HILL-ROBERTSON EN EL GENOMA

Para estimar cuantas substituciones adaptativas de cambio de aminoácido han dejado de fijarse debido al efecto Hill-Robertson hemos procedido de la siguiente forma. Siendo  $K_{a+(i)}$ ,  $L_{a(i)}$  y  $RR_{(i)}$  la tasa estimada de  $K_{a+}$ , el número total de sitios 0-veces degenerados y la tasa de recombinación promedio para el  $i$  grupo de genes (agrupados de acuerdo a su tasa de recombinación), respectivamente, hemos realizado una regresión local (LOESS, de *Local regrESSion*) entre el valor de  $K_{a+}$  y la tasa de recombinación para el  $i$  grupo de genes. El valor de la curva LOESS para cada grupo  $i$  de genes lo denominamos  $K_{a+(i)}'$  – este valor se interpreta como la tasa predicha promedio de substituciones no-sinónimas adaptativas para genes con un

valor de recombinación igual al observado. Nuestra estima de  $K_{a+}$  en ausencia de interferencia de Hill-Robertson asumimos que es el promedio de valores de  $K_{a+}$  para genes con tasas de recombinación > 2 cM/Mb y la denominamos  $K_{a+,no_{HRI}}$ . El número esperado total de substituciones no-sinónimas adaptativas en ausencia de interferencia de Hill-Robertson es, por tanto:

$$Total_{K_{a+,no_{HRI}}} = \sum(L_{a(i)} \times K_{a+,no_{HRI}}) \quad (2.15)$$

y el número de substituciones adaptativas perdidas debido a la interferencia de Hill-Robertson es:

$$Total_{K_{a+,lost}} = \sum(L_{a(i)} \times (K_{a+,no_{HRI}} - K_{a+(i)}')) \quad (2.16)$$

para aquellos grupos de genes con tasas de recombinación < 2 cM/Mb. Finalmente, la proporción de substituciones perdidas debido al efecto Hill-Robertson es:

$$f_{HRI} = \frac{Total_{K_{a+,lost}}}{Total_{K_{a+,no_{HRI}}}} \quad (2.17)$$

Como el valor promedio de substituciones no-sinónimas adaptativas de la regresión local puede ser negativo, hemos repetido el análisis cambiando todos los valores negativos de  $K_{a+(i)}'$  a cero. Una vez tenemos la  $f_{HRI}$  estimada nos preguntamos si esta es significativamente distinta de cero.

### 2.10.3 PRUEBA DE PERMUTACIÓN, BOOTSTRAPS, CÁLCULO DE INTERVALOS DE CONFIANZA Y VALORES P

Para probar que los genes del sistema inmune y de expresión sesgada en machos tenían realmente niveles de adaptación (estimada mediante  $K_{a+}$ ) mayores al resto de genes (Pröschel *et al.* 2006; Haerty *et al.* 2007; Obbard *et al.* 2009) realizamos una prueba de permutación (*permutation test*). Brevemente, este tipo de test se basa en la aleatorización de los datos y simula una distribución nula con la que comparar nuestro valor observado. Normalmente este tipo de test se aplica cuando

desconocemos la forma de la distribución nula y no podemos aplicar un test paramétrico clásico. Queremos saber si el valor de  $K_{\alpha+}$  para los genes etiquetados como genes del sistema inmune y de expresión sesgada en machos es significativamente mayor al valor de  $K_{\alpha+}$  encontrado para el resto de genes. Nuestro estadístico es entonces la diferencia entre el valor de  $K_{\alpha+}$  entre uno y otro grupo. En nuestro caso el test de permutación consiste en aleatorizar sin reemplazamiento 1.000 veces las etiquetas de los genes y estimar la diferencia entre  $K_{\alpha+}$  para los dos grupos aleatorizados de genes en cada réplica. Finalmente, calculamos el valor p para el valor observado del estadístico – el valor p es la fracción de la distribución nula que está por encima del valor observado del estadístico (o el número de réplicas con un valor del estadístico mayor al observado). El valor p es de una sola cola porque queremos probar si los genes del sistema inmune y expresión sesgada en machos tienen más cambios adaptativos que el resto de genes.  $K_{\alpha+}$  ha sido estimado mediante el software DFE-alpha para cada grupo de genes (véase sección 2.7). La distribución nula para el estadístico se puede encontrar en la figura S4 del ANEXO.

Para calcular el intervalo de confianza (IC) al 95% de la proporción de substituciones adaptativas perdidas debido a la interferencia de Hill-Robertson ( $f_{HRi}$ ) hemos realizado un remuestreo con reemplazamiento (*bootstrap*) 1.000 veces a nivel de gen. Cada uno de los 1.000 conjuntos de datos aleatorizados lo hemos dividido en 45 grupos (de 136 genes cada uno) de acuerdo a su tasa de recombinación y hemos reestimado  $K_{\alpha+}$  para cada grupo de genes independientemente utilizando el software DFE-alpha (Eyre-Walker y Keightley 2009, véase sección 2.7). Para cada conjunto de datos aleatorizado hemos realizado una regresión local entre  $K_{\alpha+}$  y la tasa de recombinación (véase sección anterior) y hemos re-estimado la  $f_{HRi}$  (expresión [2.17]).

Para probar si los genes con altas tasas de mutación (y/o elevada densidad génica) han perdido más substituciones adaptativas que los genes con bajas tasas de

mutación (y/o baja densidad génica), hemos aprovechado el *bootstrap* anterior a nivel de gen y cada una de las 1.000 réplicas la hemos dividido en dos mitades iguales, primero de acuerdo a los niveles de densidad génica y después de acuerdo a la tasa de substitución en las posiciones 4-veces degeneradas ( $K_4$ ). No obstante, antes de ordenar y categorizar los genes por sus niveles de  $K_4$  hemos generado dos estimas estadísticamente independientes de  $K_4$  (muestreando a partir de una distribución hipergeométrica, consultese sección 2.6); esto es necesario para evitar que el modo en el que hemos categorizado los genes afecte nuestros resultados. Una vez tenemos las dos estimas independientes de  $K_4$ , utilizamos  $K_{4,1}$  para definir los grupos de alta y baja tasa de mutación y  $K_{4,2}$  para estimar  $K_{\alpha+}$ . Finalmente, estimamos la  $f_{HRI}$  (expresión [2.18]) para cada réplica del *bootstrap* y cada una de las 4 categorías de genes, las cuales hemos denominado del siguiente modo: GenH-MutH (*high gene density and high mutation rate genes*), GenH-MutL (*high gene density and low mutation rate genes*), GenL-MutH (*low gene density and high mutation rate genes*), y GenL-MutL (*low gene density and low mutation rate genes*).

Con tal de probar si el  $f_{HRI}$  difiere entre las distintas categorías de genes hemos calculado el estadístico  $Z$  con la siguiente expresión:

$$Z = f_{HRI(GenH-MutH)} - f_{HRI(GenL-MutL)} \quad (2.18)$$

donde  $f_{HRI(GenH-MutH)}$  es la proporción de sustituciones adaptativas pérdidas en genes con elevadas tasas de mutación localizados en regiones de alta densidad génica y  $f_{HRI(GenL-MutL)}$  es la proporción de sustituciones adaptativas pérdidas para genes con bajas tasas de mutación en regiones de baja densidad génica. Hemos realizado todas las combinaciones posibles entre los 4 grupos de genes para obtener 6 distribuciones distintas del estadístico  $Z$ . Finalmente, para cada distribución  $Z$  hemos calculado el valor  $p$  de una sola cola contando el número de réplicas con un valor mayor (o menor) a cero. Del mismo modo, para probar si el promedio de  $K_{\alpha+}$  difiere entre categorías de genes hemos substituido la  $f_{HRI}$  por el promedio de  $K_{\alpha+}$  en la expresión (2.18).

## 2.11 SOFTWARE Y SCRIPTS

Las simulaciones analíticas han sido implementadas en un *Notebook* de *Mathematica* el cual puede consultarse en el ANEXO. Las simulaciones *forward in time* fueron ejecutadas con el software SFS\_CODE (Hernandez 2008), para extraer el espectro de frecuencias de sitios sinónimos y no-sinónimos se utilizó el programa convertSFS\_CODE (Hernandez 2008). Para generar la secuencia *re-anotada* necesaria para el análisis por ventanas no solapantes se utilizó el software desarrollado por Ràmia *et al.* (2012) (sección 2.2.1, el código en *Perl* se recoge en el ANEXO). El conteo del número de sitios, sustituciones, la corrección de Jukes y Cantor (1969) y el cálculo del SFS se realizó a través de *scripts* de *Perl* que integran tanto código genuino como partes de código procedente de: PDA2 (Casillas y Barbadilla 2006), MKT (Egea *et al.* 2008), PopDrowser (Mackay *et al.* 2012; Ràmia *et al.* 2012) y VariScan 2 (Hutter *et al.* 2006), mientras que la corrección de Tamura (1992) se realizó a parte utilizando el software MEGA-CC (Kumar *et al.* 2012) integrado en un *pipeline* de Bash (véase sección 2.2.2, todo estos *scripts* se recogen en el ANEXO). Para estimar la *DFE* de nuevas mutaciones deletéreas (Keightley y Eyre-Walker 2007) y la tasa de evolución adaptativa (Eyre-Walker y Keightley 2009) hemos utilizado el paquete de programas contenidos en la versión descargable de *DFE-alpha* (versión 2.14, disponible en URL: <http://www.homepages.ed.ac.uk/pkeightl//software>). El software CodonW (<http://codonw.sourceforge.net/>) (Peden 1999) ha sido utilizado para estimar el sesgo en el uso de codón. Creamos una lista de términos GO relacionados con la respuesta inmune, testículos y producción de esperma utilizando la herramienta QuickGO (Binns *et al.* 2009). El análisis estadístico se realizó con el paquete estadístico R (versión 3.0.2) (*R Core Team* 2013) (sección 2.10); todos los *scripts* de R están disponibles previa petición a D. Castellano (email: [castellanoed@runbox.com](mailto:castellanoed@runbox.com)). Finalmente, para realizar todos los gráficos se utilizó la biblioteca ggplot2 (Wickham 2009).

# RESULTADOS

---

En el primer bloque de resultados (sección 3.1) se han desarrollado dos estimadores puntuales que miden la acción de la selección purificadora. Estos estimadores son:  $d_n$  y  $b$ . Mediante simulaciones se ha estudiado cómo la *DFE*, el número de sitios segregantes, la selección en sitios ligados y la demografía afectan a las propiedades de dichos estimadores. En el segundo bloque (sección 3.2) se han comparado nuestros estimadores con estimadores preexistentes de la *DFE* de nuevas mutaciones deletéreas (Keightley y Eyre-Walker 2007) y se ha estimado la tasa de evolución adaptativa (Eyre-Walker y Keightley 2009) a lo largo del genoma codificador y no-codificador de *D. melanogaster*. Además, se ha estudiado el papel de la tasa de recombinación sobre la proporción de mutaciones efectivamente seleccionadas. En el tercer y último bloque (sección 3.3) se ha investigado el papel de la tasa de recombinación, la tasa de mutación y la densidad génica sobre la tasa de evolución adaptativa de mutaciones puntuales de cambio de aminoácido. Finalmente, se ha cuantificado cuantas sustituciones adaptativas de este tipo se pierden debido a la interferencia de Hill-Robertson (iHR).

## 3.1 DEFINICIÓN DE LOS NUEVOS ESTIMADORES DE LA SELECCIÓN NEGATIVA

Nuestros estimadores están inspirados en la prueba de McDonald y Kreitman (1991) y se basan el contraste del espectro de frecuencias de sitios putativamente seleccionados y sitios putativamente neutros (véase sección 1.3.2). El primero de los estimadores es el constreñimiento en la fase polimórfica para una muestra de tamaño  $n$  el cual denominamos  $d_n$ .

$$d_n = 1 - \frac{p_{sel}/L_{sel}}{(p_{neu}+1)/L_{neu}} \quad (3.1)$$

donde  $p_{sel}$  es el número de sitios segregantes y  $L_{sel}$  es el número de posiciones pertenecientes a la clase putativamente seleccionada, respectivamente.  $p_{neu}$  es el número de sitios segregantes y  $L_{neu}$  es el número de posiciones pertenecientes a la clase putativamente neutra, respectivamente. Añadimos un +1 en el denominador para evitar estimas indefinidas cuando no hay variantes neutras presentes. Las implicaciones de esta corrección han sido estudiadas en la sección 3.1.1. En el análisis de la variación del genoma codificador, los sitios selectivos suelen ser los sitios 0-veces degenerados o no-sinónimos y los sitios neutros son los sitios 4-veces degenerados o sinónimos. Aunque la selección también puede buscarse en las regiones transcritas pero no traducidas (UTRs), regiones intergénicas o intrones, y otros sitios potencialmente sometidos a selección como pequeñas secuencias reguladoras o lugares de unión a factores de transcripción. Además de los sitios sinónimos, en *D. melanogaster* también se tratan a las mutaciones puntuales que ocurren en intrones cortos < 66 pb como estándar de variación neutra (Halligan y Keightley 2006; Parsch *et al.* 2010).

El valor estimado de  $d_n$  está comprendido entre 0 y 1 y depende fuertemente de la DFE, pues la probabilidad de que una mutación seleccionada segregue respecto a una mutación neutra en una muestra de tamaño  $n$  es una función directa del coeficiente de selección de la primera. Por ejemplo, la probabilidad de que un sitio neutro contenga un polimorfismo (en el equilibrio mutación-deriva,  $P_{seg}$ ) en una muestra de tamaño  $n = 128$  es  $P_{seg} = \Theta a_n$  donde  $\Theta$  es un estimador de la variación neutra en equilibrio mutación-deriva (Tajima 1983) y  $a$  es la corrección de Watterson ( $a$  es la suma de  $1/i$  desde  $i = 1$  a  $i = n-1$ ) (Charlesworth y Charlesworth 2010; p. 29). En poblaciones no africanas de *Drosophila* donde la  $n$  sinónima es  $\sim 0,01$  esto equivale a un  $\sim 5.4\%$  para una muestra de  $n = 128$ . En cambio, la probabilidad de que un sitio selectivo contenga un polimorfismo (en el equilibrio mutación-selección-deriva) en una muestra de tamaño  $n = 128$  es  $P_{seg} = n\Theta/(4N_e s_h)$  (Loewe *et al.* 2006, expresión [8]),

donde  $s_h$  es el coeficiente de selección del heterocigoto. Una mutación con  $4N_e s_h = 100$  tiene una probabilidad de segregar del ~24% respecto a una mutación neutra en una muestra de  $n = 128$ , pero tan solo de un ~3% en una muestra de  $n = 8$ . Ambos tipos de mutación tendrán una probabilidad muy similar de segregar cuando  $n$  sea ~590. Por lo tanto, esperamos una relación negativa entre  $d_n$  y el coeficiente de selección de nuevas mutaciones deletéreas (y una relación negativa entre  $d_n$  y  $n$  para un mismo coeficiente de selección, véase sección 4.1), las propiedades de esta relación dependerán de la *DFE* para las nuevas mutaciones que ocurran en dicho gen.

Nuestro segundo estadístico estima el exceso de variantes seleccionadas a baja frecuencia respecto a las variantes neutras a baja frecuencia y lo denominamos  $b$ .

$$b = \frac{p_{sel<5\%}}{p_{sel}+1} - \frac{p_{neu<5\%}}{p_{neu}+1} \quad (3.2)$$

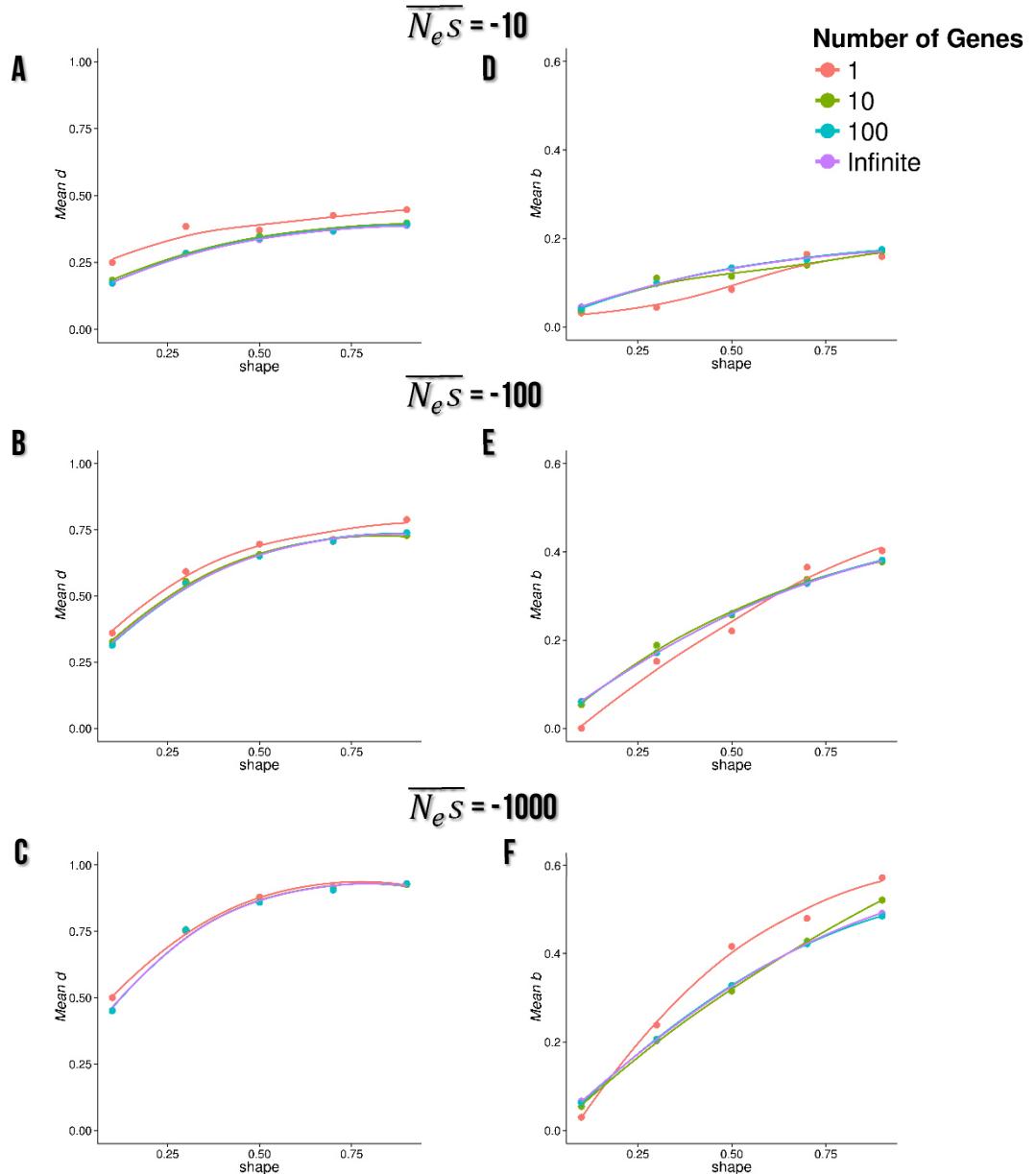
Donde  $p_{sel<5\%}$  y  $p_{neu<5\%}$  son el número de sitios segregantes de la clase seleccionada y neutra por debajo del 5%, respectivamente. No hemos observado diferencias cualitativas en las estimas si disminuimos o aumentamos hasta el 15% este valor umbral arbitrario. Podemos interpretar  $b$  intuitivamente como la proporción de sitios segregantes selectivos que podrían estar sometidos a selección purificadora directa. Es decir, si  $b = 0,2$  diremos que de todos los sitios segregantes de la clase seleccionada, un 20% de ellos segregan en exceso a baja frecuencia respecto a lo esperado para mutaciones neutras, esto es una evidencia de que estos sitios están sometidos a selección negativa (en la sección 3.1.3 se muestra cómo, dependiendo del número de sitios segregantes, este valor será o no estadísticamente significativo). El valor estimado de  $b$  está comprendido entre 0 y 1 y depende también de la *DFE*, pues el hecho que el espectro de frecuencias de mutaciones seleccionadas bajo el equilibrio mutación-selección-deriva esté enriquecido en variantes a baja frecuencia respecto al espectro esperado para mutaciones neutras es una función directa de la *DFE*.

### 3.1.1 SIMULACIONES CON SITIOS QUE SEGREGAN LIBREMENTE

Se han llevado a cabo simulaciones para investigar el comportamiento esperado de nuestros estadísticos bajo condiciones especificadas. Hemos obtenido datos mediante la aproximación de ecuaciones de difusión (Kimura 1983) donde los sitios segregan independientemente entre sí, la población está en equilibrio mutación-selección-deriva y el censo efectivo es constante (véase sección 2.3.1 para más detalles sobre las simulaciones). En estas simulaciones se han explorado 15 DFEs distintas (tabla 3.1) y se ha cuantificado el hecho que las estimas se basen en pocos sitios segregantes. No contemplamos la aparición de nuevas mutaciones beneficiosas ya que estas permanecen poco tiempo segregando en las poblaciones (figura 1.2B). El número de sitios segregantes neutros simulados para la clase neutra es variable y corresponde al valor observado en nuestra población norteamericana de *D. melanogaster* para una muestra de tamaño  $n = 128$  a nivel de 1 gen, 10 genes, 100 genes e infinitos genes. La mediana para el número de sitios segregantes sinónimos 4-veces degenerados por gen en nuestra población es de 14 y la media de 15,2 (para  $n = 128$ ). El número de sitios segregantes simulados en la clase seleccionada depende además de la DFE.

**TABLA 3.1** COMBINACIÓN DE PARÁMETROS DE LA DISTRIBUCIÓN GAMMA UTILIZADA PARA SIMULAR LA DFE

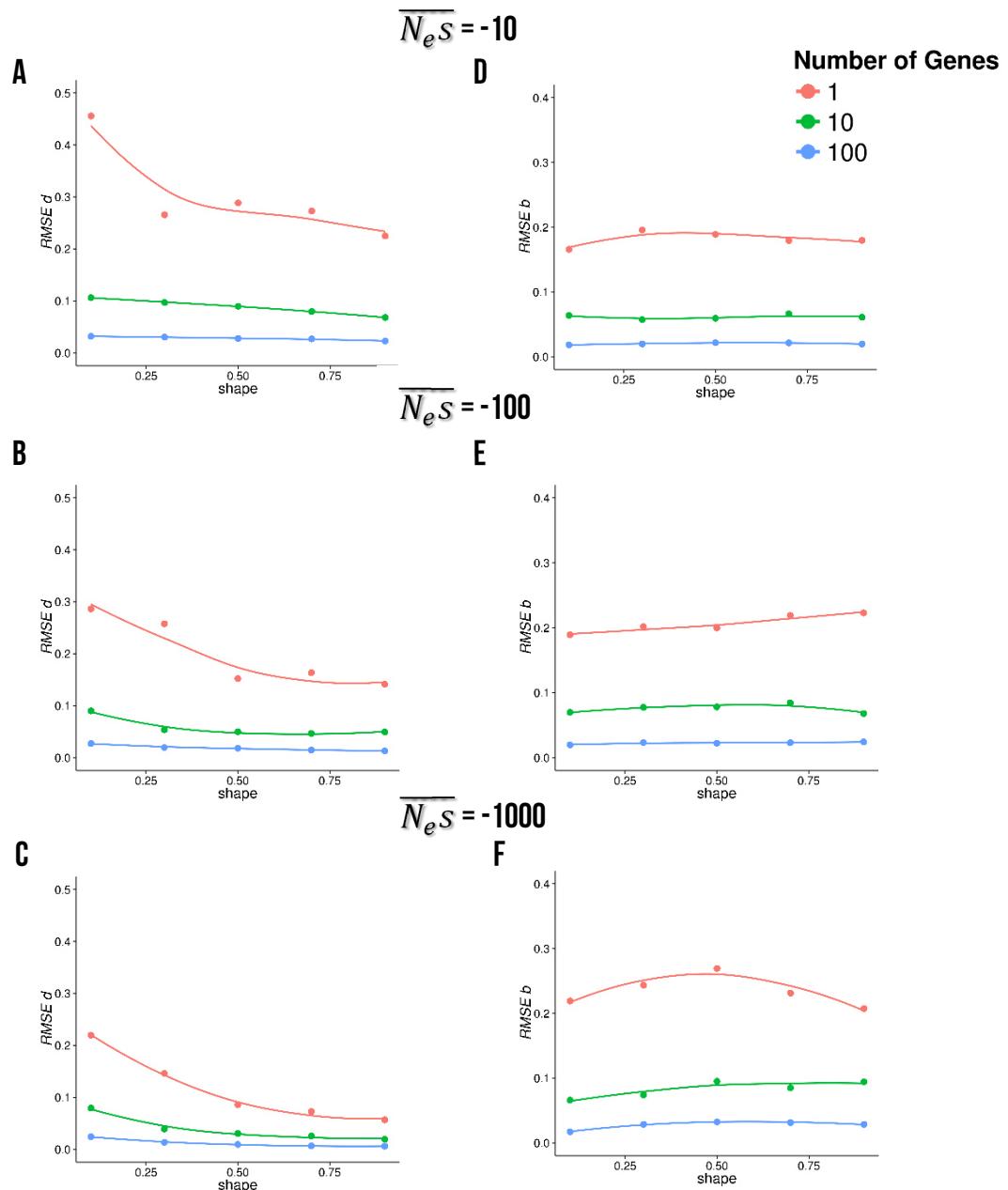
$\beta$	$\overline{N_e s}$		
0.1	-10	-100	-1000
0.3	-10	-100	-1000
0.5	-10	-100	-1000
0.7	-10	-100	-1000
0.9	-10	-100	-1000



**FIGURA 3.1** Valor promedio de los estadísticos  $d_n$  y  $b$  ante simulaciones con sitios que segregan libremente y censo efectivo constante. Cada punto es el resultado de promediar 100 iteraciones sobre la misma combinación de parámetros (tabla 3.1). Paneles A-C: valor promedio del estadístico  $d_n$ . Paneles D-F: valor promedio del estadístico  $b$ .

La figura 3.1 A-C muestra una relación positiva entre la media observada de  $d_n$  y el parámetro que determina la forma de la distribución gamma,  $\beta$  (*shape parameter*) que hemos utilizado para modelar la DFE. Esta relación es positiva para los tres efectos promedio de las nuevas mutaciones deletéreas sobre la fitness,  $|\bar{N_e S}|$  (10, 100 y 1000). Esto es así porque a medida que el parámetro de forma aumenta, la distribución se hace más platicúrtica, o uniforme, y la proporción de mutaciones seleccionadas ( $N_e s < -1$ ) respecto a las mutaciones efectivamente neutras ( $-1 < N_e s < 1$ ) aumenta. A su vez, para un mismo parámetro de forma, cuando el  $|\bar{N_e S}|$  promedio aumenta, aumenta la fuerza de la selección y  $d_n$ . Como es de esperar la fracción de mutaciones selectivas segregando en exceso a baja frecuencia respecto a las mutaciones neutras  $b$ , también aumenta con el parámetro de forma y con  $|\bar{N_e S}|$  (figura 3.1 D-F). El impacto de la corrección (+1) aplicada a los sitios segregantes en el denominador (véanse expresiones [3.1] y [3.2]) sobre las estimas de  $d_n$  y  $b$  es sólo apreciable cuando las estimas se realizan a nivel de un gen (figura 3.1). En este caso la corrección introduce un sesgo menor respecto al valor esperado con infinitos sitios.

A primera vista ambos estimadores parecen buenos estimadores de la intensidad de la selección purificadora, pues el valor promedio obtenido tras 100 réplicas con un número de genes limitado está muy cerca del valor esperado con infinitos genes. Además, el valor de ambos estadísticos aumenta al aumentar el  $|\bar{N_e S}|$  promedio de las nuevas mutaciones y/o al aumentar la proporción de mutaciones efectivamente seleccionadas ( $\beta$ ). No obstante, en la figura 3.2 observamos como el  $s_{\bar{x}}$  (que aparece como RMSE, *root mean square error*, en la figura por sus siglas en inglés), esto es la desviación estándar de las estimas respecto al valor verdadero (consúltese sección 2.3.1 y expresión [2.6] para una definición más formal), es considerablemente elevado tanto para  $d_n$  (figura 3.2 A-C) como para  $b$  (figura 3.2 D-F) a nivel de 1 gen respecto al valor esperado bajo un número infinito de sitios segregantes, pero a partir de 10-100 genes las estimas obtenidas están muy cerca del valor esperado.



**FIGURA 3.2**  $s_{\bar{x}}$  de los estadísticos  $d_n$  y  $b$  ante simulaciones con sitios que segregan libremente y censo efectivo constante. Cada punto es el resultado de estimar  $s_{\bar{x}}$  en 100 iteraciones sobre la misma combinación de parámetros (tabla 3.1). Paneles A-C:  $s_{\bar{x}}$  del estadístico  $d_n$ . Paneles D-F:  $s_{\bar{x}}$  del estadístico  $b$ .

Cuando la fuerza de la selección es baja ( $|\bar{N}_e s| = 10$ ) y la proporción de mutaciones efectivamente neutras es alta ( $\beta = 0,1$ ) las estimas de  $d_n$  son peores ( $s_{\bar{x}}$  alto) debido al aumento de la estocasticidad en el número de sitios segregantes en la clase bajo selección (figura 3.2A). En cambio, el  $s_{\bar{x}}$  de las estimas de  $b$  parece depender principalmente del número de genes y no tanto de la DFE (figura 3.2 D-F).

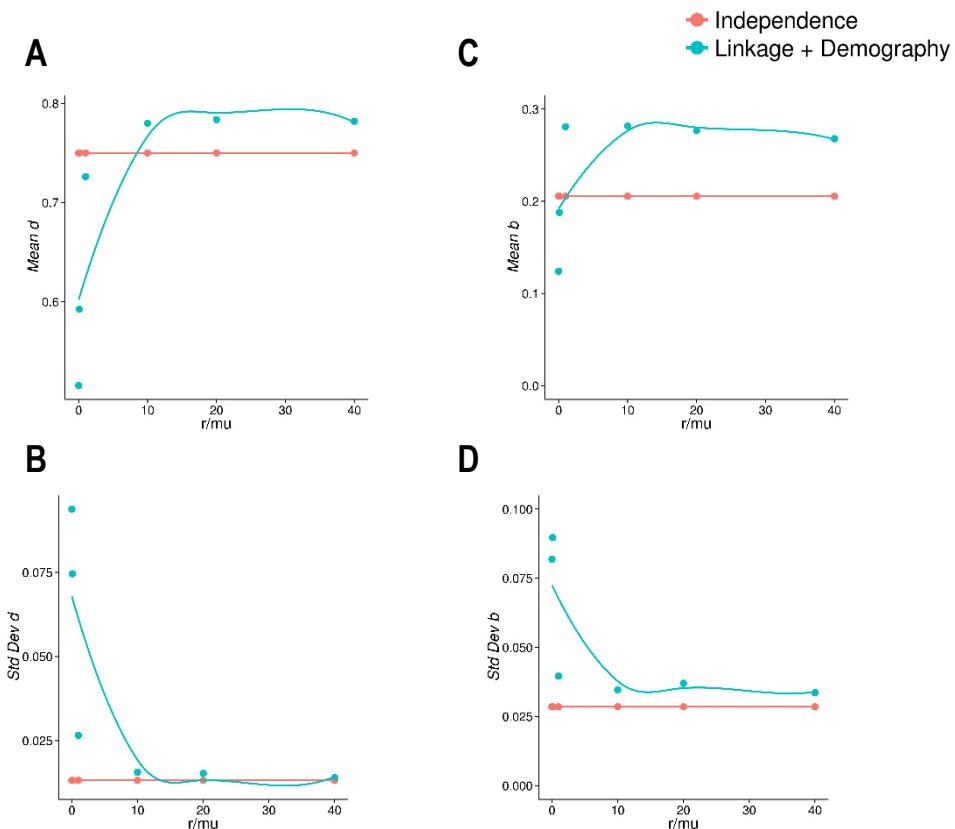
En resumen, a nivel de gen vemos mucha variación en las estimas de  $d_n$  (y  $b$ ) simplemente debido al error de muestreo. La variación en  $d_n$  (y  $b$ ) a esta escala no es variación genuina o sistemática, no es debida a que estos genes tengan distintas DFEs, es variación que proviene del error estadístico inherente a estimas basadas en tan pocos datos, o sitios segregantes. No es adecuado utilizar estos estimadores a nivel de gen, pero con tan solo 10 genes las estimas son muy precisas, y con 100 rozamos el valor verdadero esperado con infinitos sitios segregantes.

### 3.1.2 SIMULACIONES CON SITIOS LIGADOS Y DEMOGRAFÍA

El hecho que ambos estimadores relativicen la variación selectiva respecto a la variación neutra debería permitirnos controlar *a priori* los efectos de la demografía y la interferencia entre alelos sobre ambos tipos de sitios. Sin embargo, cabe la posibilidad que bajo condiciones de no-equilibrio las mutaciones neutras y selectivas muestren dinámicas poblacionales diferentes.

Para estudiar el comportamiento de nuestros estimadores en situaciones más realistas, fuera del equilibrio, estos han sido aplicados sobre datos obtenidos mediante simulaciones (*forward in time*) (Hernandez 2008) donde la DFE, la tasa de mutación, la densidad génica, la tasa de recombinación y la demografía (esto es un cuello de botella severo y reciente) se asemejan a la inferida en *D. melanogaster* en otros trabajos (véase sección 2.3.2). En esta ocasión hemos simulado una única DFE con  $|\bar{N}_e s| = 1000$  y  $\beta = 0,3$ , estos son los parámetros de la DFE para nuevas

mutaciones no-sinónimas estimados por Keightley y Eyre-Walker (2007) en *D. melanogaster*. Hemos simulado una región con 100 genes (de 1,25 kb) espaciados por 9 kb de ADN no-codificador (neutro) para 6 valores de recombinación distintos. Esto equivale a una región de poco más de una Mb. Finalmente, hemos estimado  $d_n$  y  $b$  agrupando directamente todos los sitios segregantes en los 100 genes. Consultese sección 2.3.2 para conocer los detalles de estas simulaciones.

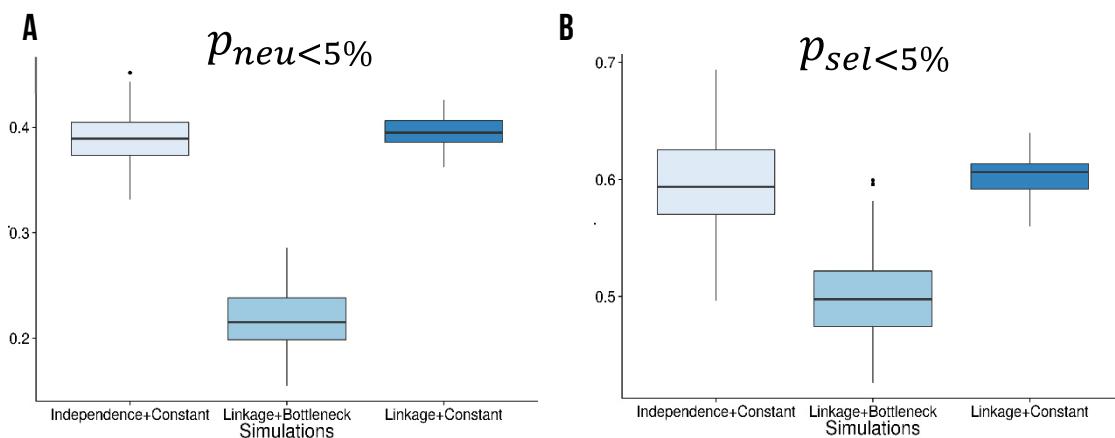


**FIGURA 3.3** Valor promedio (A y C) y desviación estándar (B y D) de las estimaciones de los estadísticos  $d_n$  y  $b$  ante simulaciones con sitios ligados en una población que ha sufrido un cuello de botella severo recientemente (puntos azules) y ante simulaciones con sitios independientes y tamaño de población constante (puntos rojos). El eje de la x muestra la razón  $r/\mu$  (consultese texto principal). Cada punto se ha estimado a partir de 100 iteraciones independientes con la misma combinación de parámetros.

La figura 3.3A muestra como la media de  $d_n$  disminuye (hasta un ~30%) respecto a los valores esperados bajo segregación independiente y censo efectivo constante a medida que la razón  $r/\mu$  se reduce. Es importante destacar que a partir de  $r/\mu > 1$  la media (y desviación estándar) bajo demografía y sitios ligados es muy similar a la encontrada bajo segregación independiente y censo efectivo constante. La media de  $d_n$  bajo demografía y sitios ligados para  $r/\mu > 1$  es tan solo un 2-3% superior a la estimada bajo independencia entre sitios y censo efectivo constante. La figura 3.3B muestra que aunque la desviación estándar de las estimas en presencia de interferencia entre sitios y demografía es bastante mayor a la esperada bajo segregación independiente, sobre todo para valores muy bajos de  $r/\mu$ , parece que sigue siendo seguro aplicar  $d_n$  a esta escala pues en el peor de los casos (cuando  $r/\mu = 0$ ) tenemos una desviación estándar en las estimas de tan solo un 9.3%.

La figura 3.3C muestra como la media de  $b$  es ~40% menor a la esperada bajo segregación independiente para  $r/\mu < 1$  pero un ~ 40% mayor a lo esperado para  $r/\mu > 1$ . La figura 3.3D muestra de nuevo que para tasas de recombinación muy bajas la desviación estándar de las estimas aumenta, pero sin anular la aplicabilidad del estadístico. La reducción de  $b$  para valores bajos de  $r/\mu$  se debe a la reducción de la eficacia de la selección purificadora en regiones de baja recombinación que hace que ambos espectros de frecuencias (selectivo y neutro) se asemejen más entre ellos. En cambio, el aumento de  $b$  para valores intermedios y altos de  $r/\mu$  puede ser debido, o bien a un incremento de la fracción de alelos selectivos segregando a baja frecuencia o bien a una reducción de la fracción de alelos neutros segregando a baja frecuencia, o a ambos procesos a la vez. La eliminación de alelos a baja frecuencia (< 5%) es esperable tras un cuello de botella severo reciente (Allendorf 1986; Denniston 1978; Nei *et al.* 1976; Maruyama y Fuerst 1985; Watterson 1984; Luikart *et al.* 1998), como el experimentado por las poblaciones no africanas de *D. melanogaster*. La pregunta es entonces si los alelos neutros se verán más afectados por el cuello de botella poblacional que los alelos selectivos, pues sólo de este modo podríamos

explicar el incremento en  $b$ . La figura 3.4 muestra la distribución de la fracción de alelos neutros y selectivos segregando a una frecuencia  $< 5\%$  para  $r/\mu = 10$  con y sin cuello de botella reciente y el caso donde los sitios segregan independientemente y el censo efectivo es constante. Según el trabajo de Thornton y Andolfatto (2006)  $r/\mu = 10$  es el valor promedio genómico para *D. melanogaster*. Podemos observar como tanto la fracción de alelos seleccionados como la fracción de alelos neutros es menor después de un cuello de botella reciente, independientemente de si los alelos segregan o no libremente. Sin embargo, la reducción es mayor para los alelos neutros, pues estos se reducen un 44% mientras los seleccionados se reducen tan solo un 16% respecto a lo esperado para una población de censo constante. Por lo tanto, la reducción de los alelos neutros a baja frecuencia producida por el cuello de botella parece ser la responsable de inflar el valor promedio de  $b$  respecto a lo esperado en ausencia de cuello de botella (tanto para  $r/\mu = 10$  como cuando los sitios segregan independientemente).

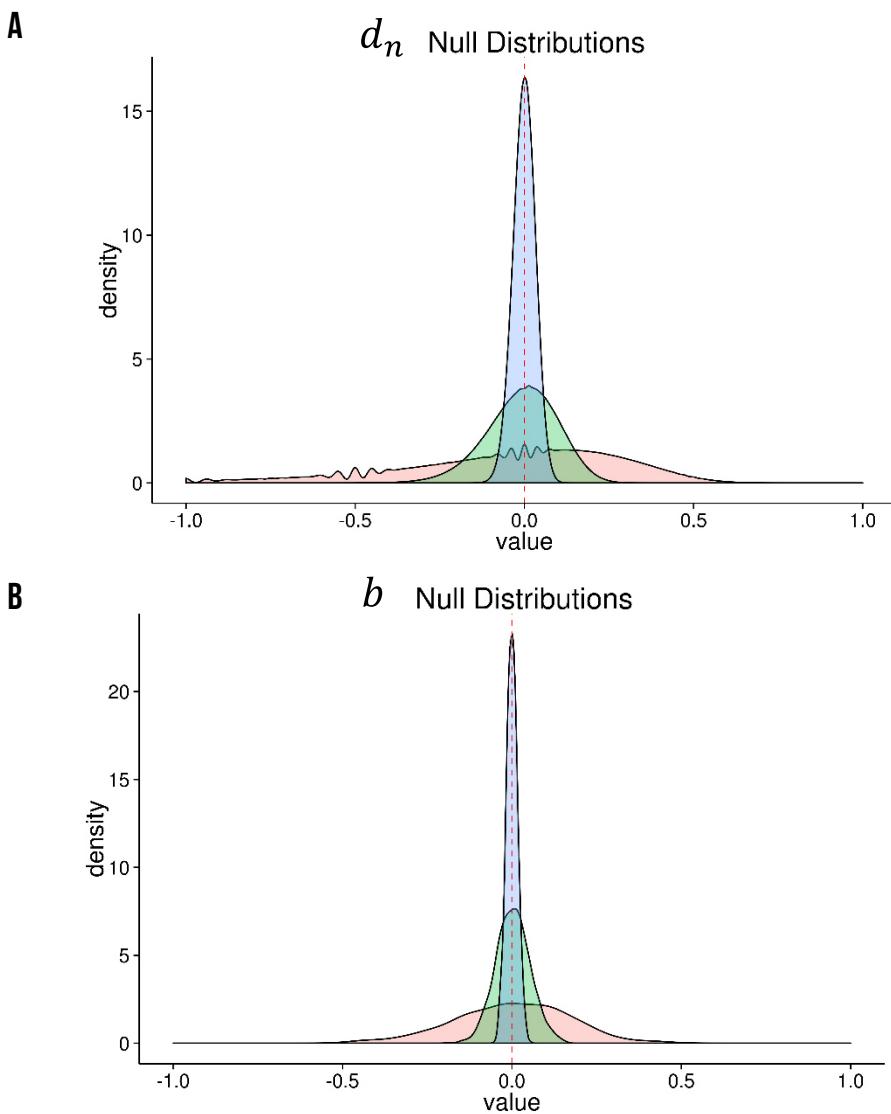


**FIGURA 3.4** Fracción de alelos neutros (A) y selectivos (B) por debajo del 5% ante distintos escenarios evolutivos. El primer diagrama de caja de ambos paneles representa la proporción de alelos  $< 5\%$  para simulaciones con sitios independientes y censo constante, el segundo diagrama de caja muestra el caso de sitios no independientes (con  $r/\mu = 10$ ) en una población que ha sufrido recientemente un cuello de botella, y el último diagrama de caja es equivalente al primero pero con sitios que no segregan independientemente (donde  $r/\mu = 10$ ). Cada diagrama de caja se ha estimado con 100 valores.

En resumen, ambos estadísticos no sólo informan sobre la *DFE* (figura 3.1), sino también informan sobre la eficacia de la selección purificadora, ya que, para valores extremadamente bajos de recombinación, tanto  $d_n$  como  $b$  disminuyen, indicando una menor eficacia de la selección purificadora debido a la disminución del  $N_e$  para mutaciones seleccionadas (efecto Hill-Robertson). En este caso sólo contemplamos la selección de fondo, nuestras simulaciones no incorporan mutaciones beneficiosas (véase sección 2.3.1). No obstante, en presencia de cuellos de botella recientes el estadístico  $b$  sobreestima la fuerza y/o la fracción de nuevas mutaciones deletéreas mientras que el estadístico  $d_n$  parece bastante robusto a la demografía (como mínimo a la demografía descrita para *D. melanogaster*) (figura 3.4).

### 3.1.3 ESTIMACIÓN DEL VALOR P DE LOS ESTIMADORES $d_n$ Y $b$

Las teorías neutralistas (Kimura 1969a; Ohta y Kimura 1971) contemplan que la gran mayoría de las nuevas mutaciones son tan deletéreas como para no segregar en la población o segregar a frecuencias bajas, y en todo caso, lo suficientemente deletéreas como para no contribuir a las diferencias entre especies. Nuestro estadístico  $d_n$  debe interpretarse como una evidencia a favor de la acción reciente de la selección purificadora fuerte y no como una refutación de la teoría neutralista (de hecho, la refuerza). A su vez, estimas significativas de  $b$  están apoyando, o demostrando, una de las predicciones más importantes de la teoría casi neutra (Ohta y Kimura 1971); la presencia de alelos ligeramente deletéreos ( $-10 < N_e s < -1$ ) segregando a baja frecuencia. Para conocer si nuestras estimas de  $d_n$  y  $b$  son significativas estadísticamente debemos contrastar el valor de los estadísticos usando datos reales con el valor de los estadísticos esperado bajo la hipótesis nula (véase sección 1.3.2). En este caso la hipótesis nula es aquella donde los sitios putativamente seleccionados son en realidad todos neutros (esto es lo que sucedería en un pseudogén por ejemplo). Dicha distribución nula debe ser estimada *ad hoc* pues depende del número de sitios segregantes implicados.



**FIGURA 3.5** Ejemplos de distribuciones nulas para los estadísticos  $d_n$  (A) y  $b$  (B) a nivel de 1 solo gen codificador en rojo, de 10 genes en verde y de 100 genes en azul. A medida que el número de datos (o genes) disponibles para realizar las estimas aumenta la distribución se estrecha, siendo más fácil obtener resultados significativos. Cada distribución contiene 10 millones de valores obtenidos tras calcular el valor de nuestros estadísticos bajo la hipótesis nula, esta es aquella donde los sitios putativamente seleccionados son todos neutros. Para hacer esto hemos generado dos variables aleatorias (muestreando a partir de una distribución de Poisson), una variable actúa como clase neutra y otra actúa como clase seleccionada. Para estimar el valor p de cada ventana y clase funcional a lo largo del genoma (véase sección 3.2) se ha realizado una distribución nula *ad hoc* de acuerdo al total de posiciones y sitios segregantes de las clases de sitios putativamente neutros y seleccionados presentes en dicha ventana.

El valor  $p$  de cada prueba corresponde a la proporción de la distribución nula que es superior (o inferior) al valor observado de  $d_n$  o  $b$ . La figura 3.5A muestra ejemplos de las distribuciones nulas que esperaríamos encontrar a nivel de 1, 10 y 100 genes (promedio) en nuestra población norteamericana de *D. melanogaster*. Cuando analizamos grupos de 100 o más genes todos los valores de  $|d_n| > |0,06|$  serán significativos (por encima, o por debajo, de  $|0,06|$  sólo están el 5% de los valores de la distribución nula), en cambio a nivel de 1 gen los valores significativos serán sólo aquellos con  $|d| > |0,5|$ . En este trabajo hemos calculado  $d_n$  y  $b$  en ventanas no solapantes de 1 Mb a lo largo del genoma de *D. melanogaster* (donde en promedio hay  $\sim 110$  genes codificadores y mucha más secuencia no-codificadora asociada) (véase siguiente sección). Para cada ventana y clase funcional putativamente seleccionada (región intergénica, UTR, intrón y posiciones 0-veces degeneradas) hemos estimado la distribución nula del estadístico  $d_n$  y  $b$  de acuerdo al número de sitios segregantes 4-veces degenerados, el total de posiciones 4-veces degeneradas y el total de posiciones putativamente seleccionadas presentes en dicha ventana (figura 3.5 para más detalles).

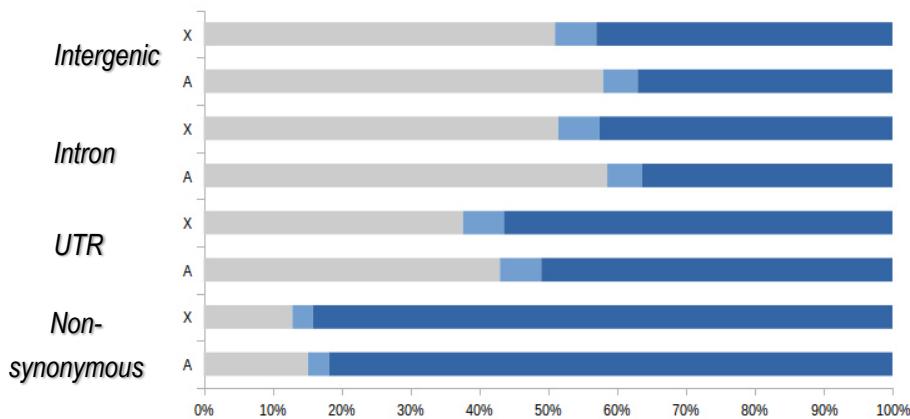
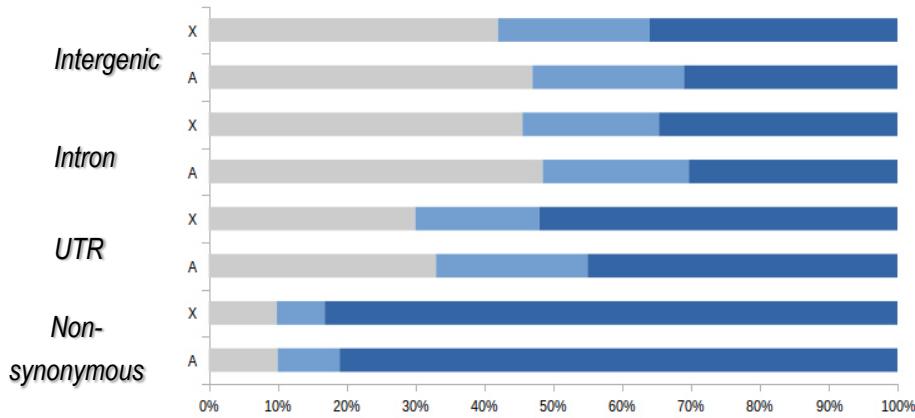
## 3.2 ESTIMACIÓN DE LA HUELLA DE LA SELECCIÓN NATURAL A LO LARGO DEL GENOMA CODIFICADOR Y NO-CODIFICADOR DE *Drosophila melanogaster*

Los estadísticos  $d_n$  y  $b$  fueron aplicados en el artículo original del proyecto DGRP (Mackay *et al.* 2012) sobre una población norteamericana de *D. melanogaster*. En análisis posteriores (no publicados) se han aplicado los estimadores existentes de la *DFE* (Keightley y Eyre-Walker 2007) sobre el mismo conjunto de datos y comparado los resultados con los resultados obtenidos mediante nuestros estimadores. La ventaja añadida de los métodos que estiman la *DFE* cuantitativamente respecto a nuestros estimadores puntuales es que estos pueden corregir mejor el efecto de las mutaciones ligeramente deletéreas sobre las estimas de la adaptación (Eyre-Walker y Keightley 2009). Todas las estimas en esta sección han sido realizadas sobre

ventanas cromosómicas no solapantes de 1 Mb (en promedio cada Mb eucromática contiene unos 110 genes – estamos dentro de la zona de confianza de nuestros estimadores). La tabla 3.2 muestra el número de bases analizadas para cada clase funcional: región intergénica, intrones, UTRs y sitios 0-veces degenerados, en autosomas y el cromosoma X por separado. Los detalles de los criterios para incluir, excluir y anotar cada base del genoma se pueden consultarse en la sección 2.2. Tanto las regiones 5' como 3' de las regiones intergénicas y UTRs han sido agrupadas bajo una misma etiqueta. Es decir, no se ha distinguido entre las regiones aguas arriba y aguas abajo de los genes. Además, dentro de las regiones intergénicas sólo se han analizado las bases 5 kb adyacentes (como máximo) a la última posición y la primera posición del último y primer exón de un mismo gen, respectivamente. Las secuencias intrónicas corresponden a las secuencias de aquellos intrones sin otros genes anidados dentro y para estos no se ha aplicado ningún límite de longitud. Finalmente, en todos los análisis que se muestran a continuación se ha asumido que los sitios sinónimos son neutros – estos han servido de secuencia de referencia neutra para encontrar evidencias de selección en otras clases de sitios putativamente seleccionados.

### 3.2.1 MUTACIONES DELETÉREAS Y EFECTIVAMENTE NEUTRAS

La figura 3.6A muestra gráficamente los valores promedio de  $d_n$  y  $b$  para los autosomas y el cromosoma X de *D. melanogaster* tanto para el ADN no-codificador (región intergénica, intrones y UTRs) como para el ADN codificador (sitios 0-veces degenerados). En esta figura se representa  $b$  como una fracción respecto al total de sitios en la clase seleccionada y no como una fracción de sitios segregantes en la clase seleccionada (véase ecuación [3.2]), esto permite representar tanto  $d_n$  como  $b$  a nivel de sitio selectivo y dar una visión escalada de cada estadístico. Asumimos que la fracción de sitios que no pertenecen ni a  $d_n$  ni a  $b$  son neutros, a esta fracción de sitios selectivos potencialmente neutros la denominamos como la fracción  $f$  (donde  $f = 1 - b - d_n$ ).

**A****B**

**FIGURA 3.6** Representación gráfica de la huella de la selección purificadora sobre secuencias codificadoras y no-codificadoras del cromosoma X y autosomas en ventanas no solapantes de 1 Mb utilizando: (A) nuestros estimadores puntuales de la intensidad y eficacia de la selección purificadora:  $f$  en gris,  $b$  en azul claro y  $d_n$  en azul oscuro (consúltense descripción en el texto principal) y (B) el método de Keightley y Eyre-Walker (2007) donde la DFE de nuevas mutaciones es estimada por el método de máxima verosimilitud. La fracción de nuevas mutaciones efectivamente neutras se representa en gris, la ligeramente deletérea en azul claro y la fuertemente deletérea en azul oscuro.

La figura 3.6B es equivalente a la figura 3.6A pero utilizando el método Keightley y Eyre-Walker (2007) (consúltese sección 2.7) para estimar la DFE de nuevas mutaciones deletéreas, donde representamos la fracción de mutaciones efectivamente neutras ( $-1 < N_{e}s < 1$ ), ligeramente deletéreas ( $-10 < N_{e}s < -1$ ) y fuertemente deletéreas ( $N_{e}s < -10$ ). El valor promedio de  $d_n$ , ponderado por la

longitud de las secuencias a lo largo de todo el genoma (codificador y no codificador, X y autosomas), es del ~47%, el de  $b$  es del ~5% y el resto  $f$  ~48% se asume evoluciona de forma neutra. Mediante el estimador cuantitativo de la *DFE* de Keightley y Eyre-Walker (2007) obtenemos que a lo largo del genoma un ~39% de las nuevas mutaciones son fuertemente deletéreas, un ~21% ligeramente deletéreas y el resto ~40% son efectivamente neutras ponderando de nuevo por la contribución al genoma de cada clase funcional.

Nuestros estadísticos son estimadores puntuales de la eficacia y fuerza de la selección purificadora, no son estimadores cuantitativos de la *DFE* aunque sí señalan importantes aspectos cualitativos de esta (véase sección 3.1.2 y figura 3.1). Realmente no sabemos el rango de coeficientes de selección representado en cada uno de nuestros estimadores, pero es lógico interpretar  $d_n$  como la fracción de mutaciones deletéreas más fuertemente seleccionadas,  $b$  como la fracción de mutaciones deletéreas más débilmente seleccionadas y  $f$  como la fracción de mutaciones muy débilmente o nada seleccionadas. En este sentido nuestros estimadores sobreestiman, en comparación al método de Keightley y Eyre-Walker (2007), la fracción de mutaciones efectivamente neutras, subestiman la fracción de mutaciones ligeramente deletéreas y sobreestiman ligeramente la fracción de mutaciones fuertemente deletéreas. A pesar de estas diferencias, la correlación entre  $d_n$  y la fracción de mutaciones fuertemente deletéreas es muy alta (correlación de Spearman  $\rho_s = 0,99$  y  $P < 0,001$ ), al igual que la correlación entre la fracción de mutaciones ligeramente deletéreas y las que segregan en exceso por debajo del 5% ( $b$ ) ( $\rho_s = 0,76$  y  $P < 0,001$ ) y la correlación entre la fracción de mutaciones efectivamente neutras y nuestro estadístico  $f$  ( $\rho_s = 0,91$  y  $P < 0,001$ ). Ambas metodologías a pesar de basarse en aproximaciones distintas, parecen dibujar prácticamente el mismo paisaje.

**TABLA 3.2 DATOS ANALIZADOS, VALORES PROMEDIO Y DESVIACIÓN ESTÁNDAR DE DISTINTOS ESTADÍSTICOS POR CLASE FUNCIONAL Y TIPO DE CROMOSOMA**

		L1	L2	$d_{sel}/d_{neu}$	$\alpha$	$\omega_A$	$-1 < N_e S < 1$	$-10 < N_e S < -1$	$N_e S < -10$	f	b	$d_n$
Non-synonymous	A	9869434	9840163	0.15(0.04)	0.40(0.38)	0.07(0.06)	0.10(0.06)	0.09(0.03)	0.81(0.07)	0.15(0.07)	0.03(0.01)	0.81(0.07)
	X	1910404	1900799	0.18(0.05)	0.53(0.37)	0.10(0.07)	0.10(0.07)	0.07(0.02)	0.84(0.06)	0.13(0.06)	0.03(0.01)	0.85(0.06)
UTR	A	3937439	3853431	0.44(0.10)	0.32(0.31)	0.14(0.15)	0.33(0.17)	0.22(0.17)	0.45(0.17)	0.43(0.15)	0.06(0.03)	0.51(0.15)
	X	773820	750234	0.51(0.07)	0.45(0.30)	0.24(0.16)	0.30(0.15)	0.18(0.10)	0.52(0.12)	0.38(0.12)	0.06(0.02)	0.57(0.11)
Intron	A	21682140	21026273	0.56(0.16)	0.20(0.34)	0.12(0.22)	0.48(0.21)	0.21(0.15)	0.30(0.16)	0.58(0.18)	0.05(0.03)	0.36(0.17)
	X	4057438	3897220	0.63(0.10)	0.34(0.29)	0.22(0.18)	0.46(0.18)	0.20(0.12)	0.35(0.13)	0.52(0.13)	0.06(0.04)	0.43(0.11)
Intergenic	A	24950212	24224311	0.53(0.16)	0.17(0.37)	0.10(0.20)	0.47(0.21)	0.22(0.14)	0.31(0.17)	0.58(0.20)	0.05(0.03)	0.37(0.19)
	X	4990937	4795947	0.62(0.15)	0.35(0.31)	0.23(0.23)	0.42(0.20)	0.22(0.17)	0.36(0.13)	0.51(0.15)	0.06(0.03)	0.43(0.12)

La primera columna muestra la clase funcional analizada. A y X corresponde a los resultados en autosomas y el cromosoma X, respectivamente. L1 es el número de bases analizadas en *D. melanogaster* y L2 es el número de bases analizadas entre *D. melanogaster* - *D. yakuba*. La diferencia entre ambos (L1-L2) es el número de gaps en el alineamiento.  $d_{sel}/d_{neu}$  es la razón entre la tasa de sustitución en sitios potencialmente funcionales y la tasa de sustitución neutra (sinónima en nuestro caso).  $d_{sel}$  ha sido corregida mediante el método de Jukes y Cantor, JC (1968), y  $d_{neu}$  mediante el método de Tamura (1992).  $\alpha$  es la fracción de substituciones adaptativas.  $\omega_A$  es la tasa de sustitución adaptativa entre la tasa de sustitución neutra.  $-1 < N_e S < 1$  es la fracción de nuevas mutaciones efectivamente neutras,  $-10 < N_e S < -1$  es la fracción de nuevas mutaciones ligeramente deletéreas y  $N_e S < -10$  es la fracción de nuevas mutaciones fuertemente deletéreas. Las últimas tres columnas corresponden al valor de nuestros estadísticos f, b y  $d_n$  (consultese texto principal para las definiciones). Los decimales se separan con puntos en lugar de comas pues toda la tabla está en inglés.

La tabla S7 del ANEXO contiene el valor de todos estos estadísticos para cada una de las ventanas y clases funcionales junto con el valor p de nuestros estimadores y la diferencia del logaritmo de la función de verosimilitud (log L) entre el modelo con selección respecto el modelo sin selección para los estimadores de Keightley y Eyre-Walker (2007). En todos los casos las pruebas son significativas indicando la omnipresencia de la selección purificadora en *D. melanogaster*.

**TABLA 3.3 DATOS ANALIZADOS, VALORES PROMEDIO Y DESVIACIÓN ESTÁNDAR DE DISTINTOS ESTADÍSTICOS POR BRAZO CROMOSÓMICO**

	L1	L2	d <sub>N</sub> /d <sub>S</sub>	α	ω <sub>A</sub>	-1 < N <sub>eS</sub> < 1	-10 < N <sub>eS</sub> < -1	N <sub>eS</sub> < -10	f	b	d <sub>n</sub>
2L	13800174	13448604	0.47(0.10)	0.34(0.21)	0.16(0.10)	0.35(0.13)	0.24(0.11)	0.41(0.13)	0.45(0.12)	0.06(0.02)	0.49(0.11)
2R	13098515	12770092	0.49(0.18)	0.28(0.23)	0.13(0.11)	0.39(0.16)	0.20(0.07)	0.41(0.14)	0.48(0.17)	0.05(0.02)	0.46(0.16)
3L	15309912	14932570	0.47(0.15)	0.20(0.38)	0.09(0.20)	0.41(0.20)	0.19(0.12)	0.40(0.16)	0.51(0.22)	0.05(0.04)	0.44(0.20)
3R	18230624	17792912	0.44(0.06)	0.03(0.38)	0.02(0.17)	0.46(0.17)	0.16(0.10)	0.38(0.14)	0.53(0.16)	0.04(0.02)	0.42(0.15)
X	11732599	11344200	0.54(0.09)	0.37(0.28)	0.20(0.16)	0.37(0.16)	0.19(0.10)	0.45(0.12)	0.44(0.13)	0.05(0.03)	0.51(0.11)

La primera columna muestra la clase funcional analizada. L1 es el número de bases analizadas en *D. melanogaster* y L2 es el número de bases analizadas entre *D. melanogaster* - *D. yakuba*. La diferencia entre ambos (L1-L2) es el número de gaps en el alineamiento. d<sub>sel</sub>/d<sub>neu</sub> es la razón entre la tasa de sustitución en sitios potencialmente funcionales y la tasa de sustitución neutra (sinónima en nuestro caso). d<sub>sel</sub> ha sido corregida mediante el método de Jukes y Cantor, JC (1968), y d<sub>neu</sub> mediante el método de Tamura (1992). α es la fracción de substituciones adaptativas. ω<sub>A</sub> es la tasa de sustitución adaptativa entre la tasa de sustitución neutra. -1 < N<sub>eS</sub> < 1 es la fracción de nuevas mutaciones efectivamente neutras, -10 < N<sub>eS</sub> < -1 es la fracción de nuevas mutaciones ligeramente deletéreas y N<sub>eS</sub> < -10 es la fracción de nuevas mutaciones fuertemente deletéreas. Las últimas tres columnas corresponden al valor de nuestros estadísticos f, b y d<sub>n</sub> (consultese texto principal para las definiciones). Los decimales se separan con puntos en lugar de comas pues toda la tabla está en inglés.

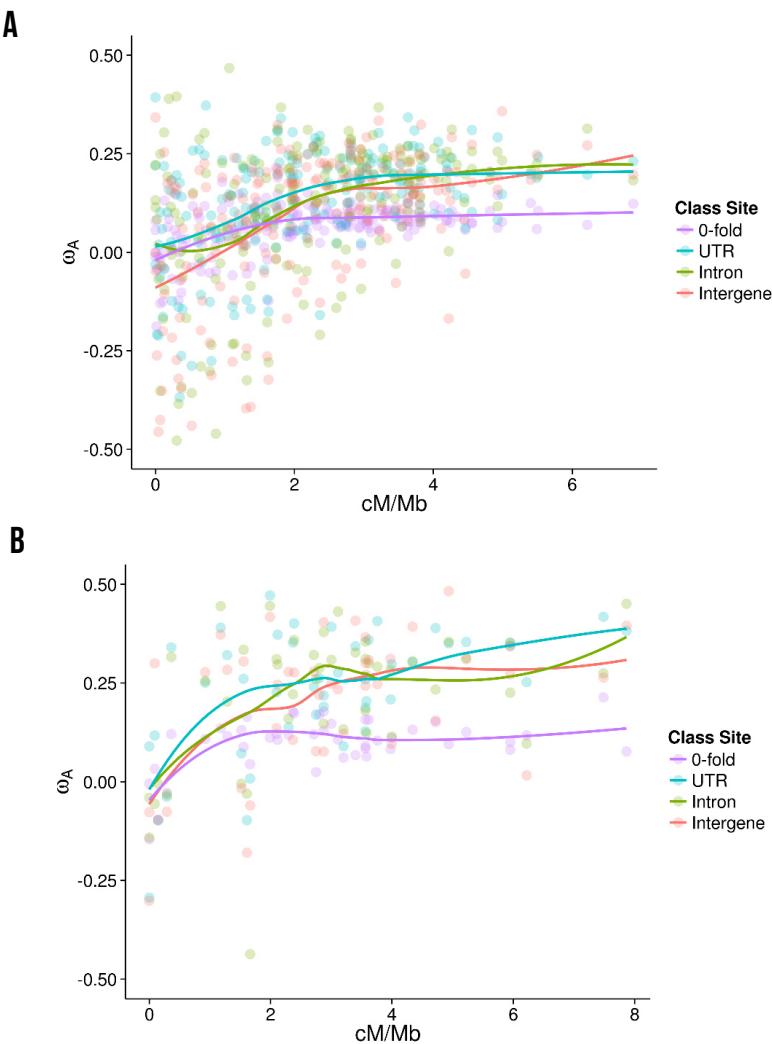
En la tabla 3.2 se muestra el valor promedio (y la desviación estándar) de la proporción de nuevas mutaciones deletéreas en cada uno de los tres intervalos de efectos sobre la *fitness* (Keightley y Eyre-Walker 2007) junto con los valores de nuestros estimadores puntuales de la eficacia e intensidad de la selección purificadora para las distintas clases funcionales. Las secuencias codificadoras son las más constreñidas de acuerdo a los valores significativamente superiores de la fracción de nuevas mutaciones fuertemente deletéreas en comparación con el ADN no-codificador (Test de Mann–Whitney  $P < 0,001$ ). Dentro del ADN no-codificador las regiones UTR son las más constreñidas (UTR vs intrones,  $P < 0,001$ ; UTR vs región intergénica,  $P < 0,001$ ). No hay diferencias significativas para la fracción de mutaciones fuertemente deletéreas entre intrones y regiones intergénicas ( $P = 0,56$ ). El ADN no-codificador muestra en conjunto una mayor proporción de mutaciones ligeramente deletéreas que las regiones codificadoras ( $P < 0,001$ ) y obviamente una mayor fracción de mutaciones efectivamente neutras ( $P < 0,001$ ). Sin embargo, la fracción de mutaciones ligeramente deletéreas es similar entre las distintas clases de ADN no-codificador (UTR vs intrones,  $P = 1$ ; UTR vs región intergénica,  $P = 0,29$ ; región intergénica vs intrones,  $P = 1$ ). La tabla 3.3 muestra el valor de nuestros estimadores y el de Keightley y Eyre-Walker (2007) para cada brazo cromosómico ponderando por la longitud de cada clase funcional. Globalmente, hay diferencias significativas en la fracción de nuevas mutaciones fuertemente deletéreas entre el brazo cromosómico 3R y el cromosoma X ( $P < 0,05$ ), mientras el resto de cromosomas no muestran diferencias significativas entre ellos. Respecto a la fracción de mutaciones ligeramente deletéreas, no hay diferencias significativas entre brazos cromosómicos. Finalmente, el brazo cromosómico 3R tiene una fracción de nuevas mutaciones efectivamente neutras significativamente mayor al brazo cromosómico 2L y el cromosoma X (2L vs 3R,  $P < 0,05$ ; X vs 3R,  $P < 0,05$ ). Estas comparaciones entre X y autosomas no tienen en cuenta dos particularidades relevantes del cromosoma X: (1) su mayor tasa de entrecruzamiento y (2) su menor densidad génica. En la sección 3.2.4 se evalúa el impacto de estas variables sobre las diferencias en la DFE.

### 3.2.2 MUTACIONES BENEFICIOSAS

La tabla 3.2 muestra el promedio del número de substituciones adaptativas por substitución neutra ( $\omega_A$ ) estimadas mediante el método de Eyre-Walker y Keightley (2009) para autosomas y X y las distintas clases funcionales.  $\omega_A$  es un estadístico más adecuado que la tradicional fracción de substituciones adaptativas ( $\alpha$ ) porque no depende de las substituciones ligeramente deletéreas o neutras, y además corrige para diferencias en la tasa de mutación entre *loci* o regiones. A nivel genómico las mutaciones no-sinónimas muestran una menor tasa de evolución adaptativa en relación a las mutaciones que ocurren en el ADN no-codificador (codificador vs UTR,  $P < 0,001$ ; codificador vs intrón,  $P < 0,001$ ; codificador vs región intergénica,  $P < 0,001$ ). La tasa de evolución adaptativa en el ADN no-codificador es en promedio ~2 veces mayor que en el ADN codificador ( $P < 0,001$ ) (tabla 3.2). No hay diferencias significativas entre tasas de adaptación para los distintos tipos de ADN no-codificador (intrón vs UTR,  $P = 1$ ; intrón vs región intergénica,  $P = 1$ ; UTR vs región intergénica,  $P = 0,19$ ) (tabla 3.2). Dentro de los autosomas el brazo cromosómico 2L muestra ~2 veces más adaptación que el resto de autosomas juntos (2L vs 2R,  $P < 0,05$ ; 2L vs 3L,  $P < 0,001$ ; 2L vs 3R,  $P < 0,001$ ) y el brazo cromosómico 3R muestra una tasa de adaptación muy inferior al resto de autosomas (3R vs 2R,  $P < 0,001$ ; 3R vs 3L,  $P < 0,001$ ) (tabla 3.3). De nuevo, las diferencias en la tasa de adaptación entre X y autosomas se estudian en detalle en la sección 3.2.4.

### 3.2.3 MUTACIONES EFECTIVAMENTE SELECCIONADAS Y RECOMBINACIÓN

El efecto Hill-Robertson (Hill y Robertson 1966) (sección 1.1.3) produce una reducción del censo efectivo ( $N_e$ ) en regiones del genoma con baja recombinación generando una relación positiva entre la tasa de adaptación y la tasa de recombinación y una relación negativa entre fracción de mutaciones efectivamente neutras y la tasa de recombinación (esto lo esperamos para todos los sitios seleccionados, tanto codificadores como no-codificadores).



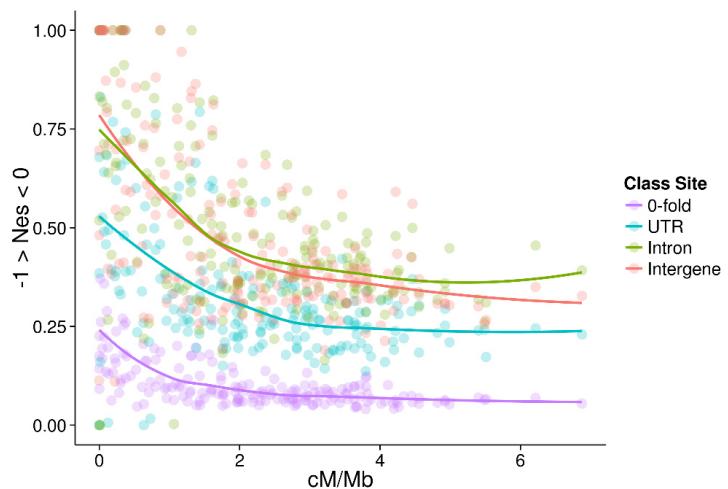
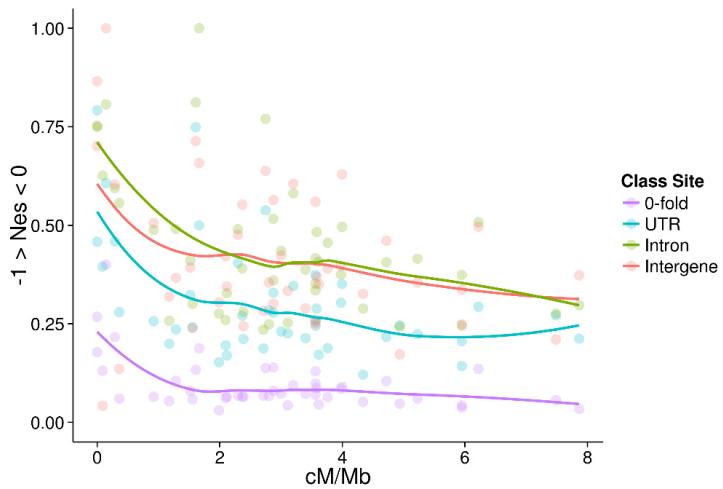
**FIGURA 3.7** Relación entre la tasa de recombinación (cM/Mb) y la tasa de evolución adaptativa ( $\omega_A$ ) en ventanas no solapantes de 1 Mb para cada clase funcional en (A) autosomas y (B) el cromosoma X. Las líneas son regresiones locales (LOESS).

La figura 3.7 muestra la relación entre  $\omega_A$  y la tasa de recombinación para cada clase funcional en los autosomas y el cromosoma X, respectivamente. Hay una correlación significativa y positiva entre  $\omega_A$  y la tasa de recombinación para todas las clases de sitios en autosomas (codificador  $\rho_s = 0,45, P < 0,001$ ; UTR  $\rho_s = 0,33, P < 0,001$ , intrones  $\rho_s = 0,38, P < 0,001$ ; región intergénica  $\rho_s = 0,36, P < 0,001$ ). En el cromosoma X la tasa de adaptación en las regiones codificadoras no está significativamente

correlacionada con la tasa de recombinación, pero sí lo está para todas las regiones no-codificadoras (codificador  $\rho_s = 0,20, P = 0,19$ ; UTR  $\rho_s = 0,48, P < 0,001$ , intrones  $\rho_s = 0,32, P < 0,05$ ; región intergénica  $\rho_s = 0,31, P < 0,05$ ). Esta falta de correlación entre adaptación proteica y recombinación en el X había sido previamente descrita (Campos *et al.* 2014) y podría ser debida a la menor densidad génica del cromosoma y/o la mayor tasa de entrecruzamientos en el X respecto a los autosomas; ambos hechos disminuirían la intensidad de la interferencia de Hill-Robertson en el X (a continuación se estudia en profundidad este resultado).

La figura 3.8 muestra la relación entre la fracción de mutaciones efectivamente neutras (estimada mediante el método de Keightley y Eyre-Walker 2007) y la tasa de recombinación para cada clase funcional de nuevo diferenciando entre autosomas y el cromosoma X. Hay una correlación significativa y negativa entre la fracción de nuevas mutaciones efectivamente neutras y la tasa de recombinación para todas las clases de sitios en autosomas (codificador  $\rho_s = -0,61, P < 0,001$ ; UTR  $\rho_s = -0,48, P < 0,001$ , intrones  $\rho_s = -0,51, P < 0,001$ ; región intergénica  $\rho_s = -0,53, P < 0,001$ ) y el cromosoma X (codificador  $\rho_s = -0,40, P < 0,01$ ; UTR  $\rho_s = -0,39, P < 0,01$ , intrones  $\rho_s = -0,40, P < 0,01$ ), a excepción de la región intergénica donde la correlación es negativa pero no significativa ( $\rho_s = -0,24, P = 0,12$ ).

En resumen, los resultados mostrados en las figuras 3.7 y 3.8 apoyan la hipótesis que dentro de un mismo genoma el  $N_e$  varía de acuerdo a la tasa de recombinación. En la sección 3.3 se profundiza en esta hipótesis incorporando la variación a lo largo de los cromosomas de la tasa de mutación y de la densidad génica.

**A****B**

**FIGURA 3.8** Relación entre la tasa de recombinación (cM/Mb) y la fracción de nuevas mutaciones efectivamente neutras ( $-1 < N_{eS} < 1$ ) en ventanas no solapantes de 1 Mb para cada clase funcional en (A) autosomas y (B) el cromosoma X. Las líneas son regresiones locales (LOESS).

### 3.2.4 CROMOSOMA X vs AUTOSOMAS

Las comparaciones de la DFE de nuevas mutaciones deletéreas y la tasa de evolución adaptativa mostradas anteriormente entre X y autosomas no tenían en cuenta la mayor tasa de entrecruzamiento en el cromosoma X respecto a los autosomas. La

recombinación es más elevada en el cromosoma X porque los machos de *Drosophila* no pueden recombinar entre cromosomas homólogos (Ashburner *et al.* 2005); la tasa de recombinación efectiva para una tasa de recombinación  $r$  dada entre dos *loci* en hembras es  $1/2$  para los autosomas y  $2/3$  para el X (Charlesworth y Charlesworth 2010, p.381). A su vez, la densidad génica del cromosoma X es un 20% menor a la de los autosomas ( $P < 0,05$ ). Ambas características tenderán a disminuir la intensidad de la interferencia de Hill-Robertson (Hill y Robertson 1966) en el X respecto a los autosomas dado que la tasa de mutación es equivalente en el cromosoma X y en los autosomas (Bauer y Aquadro 1997; Hutter *et al.* 2007; Keightley *et al.* 2009; Zeng y Charlesworth 2010; Haddrill *et al.* 2011; Hu *et al.* 2013).

**TABLA 3.4** RAZÓN X/A PARA TRES INTERVALOS DE LA DIFERENCIA  $\omega_A$  ENTRE REGIONES CROMOSÓMICAS CON UNA TASA DE RECOMBINACIÓN Y DENSIDAD GÉNICA EQUIVALENTE

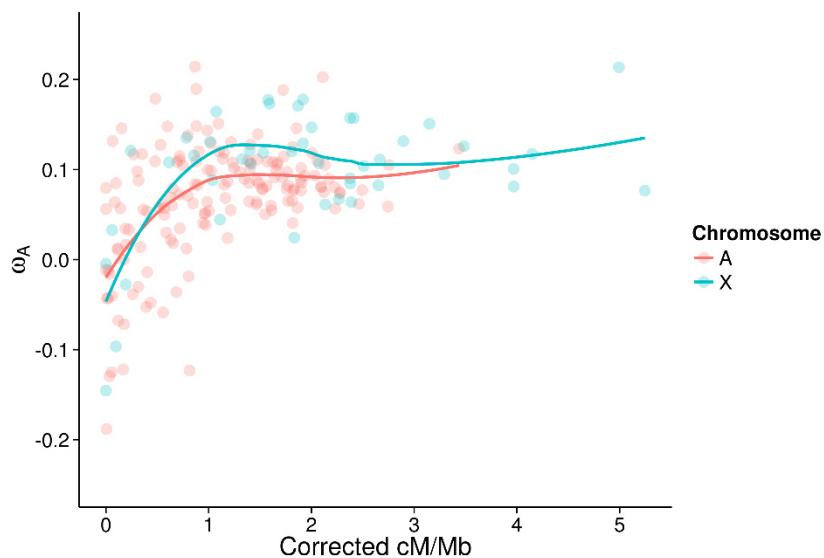
		X/A	P
Non-synonymous	-1 < $N_{eS}$ < 1	1.07	9.2e-1
	-10 < $N_{eS} < -1$	0.83	3.3e-6 ***
	$N_{eS} < -10$	1.01	1.4e-1
	$\omega_A$	1.30	2.4e-4 ***
UTR	-1 < $N_{eS} < 1$	1.06	9.3e-1
	-10 < $N_{eS} < -1$	0.98	5.9e-1
	$N_{eS} < -10$	0.98	7.5e-1
	$\omega_A$	1.45	3.4e-5 ***
Intron	-1 < $N_{eS} < 1$	1.03	5.2e-1
	-10 < $N_{eS} < -1$	0.98	4.6e-1
	$N_{eS} < -10$	0.98	5.2e-1
	$\omega_A$	1.43	6.0e-5 ***
Intergenic	-1 < $N_{eS} < 1$	1.06	7.7e-1
	-10 < $N_{eS} < -1$	0.92	2.6e-1
	$N_{eS} < -10$	0.99	6.1e-1
	$\omega_A$	1.47	3.5e-4 ***

Si comparamos el cromosoma X con cada uno de los brazos autosómicos observamos que: (1) el cromosoma X tiene una mayor tasa de adaptación global que cada brazo autosómico (sin distinguir ADN codificador y no-codificador) (X vs 2L,  $P < 0,05$ ; X vs 2R,  $P < 0,001$ ; X vs 3L,  $P < 0,001$ ; X vs 3R,  $P < 0,001$ ), (2) la fracción de mutaciones

efectivamente neutras es significativamente mayor en el brazo 3R que en el cromosoma X ( $P < 0,05$ ) y (3) la fracción de mutaciones fuertemente deletéreas es significativamente menor en el brazo 3R que en el cromosoma X ( $P < 0,05$ ) (tabla 3.3). No obstante, como se acaba de comentar, estas diferencias pueden ser debidas a la mayor tasa de entrecruzamiento en el X y/o su menor densidad génica. Si comparamos ventanas autosómicas y del X con una tasa de recombinación efectiva (estimas corregidas  $> 1 \text{ cM/Mb}$ ) y una densidad génica similar encontramos que para las mutaciones no-sinónimas no hay diferencias significativas ni para la fracción de mutaciones efectivamente neutras ( $P = 0,92$ ) ni fuertemente deletéreas ( $P = 0,14$ ), pero sí para la tasa de evolución adaptativa ( $P < 0,001$ ) y la fracción de mutaciones ligeramente deletéreas ( $P < 0,001$ ) (tabla 3.4). No hay diferencias significativas para la DFE de nuevas mutaciones deletéreas entre brazos autosómicos, ni diferencias significativas entre cada uno de los brazos autosómicos y el cromosoma X para la fracción de nuevas mutaciones efectivamente neutras y fuertemente deletéreas. La tasa de evolución adaptativa no-sinónima es 1,3 veces mayor en el cromosoma X que en el conjunto de autosomas (tabla 3.4) (la diferencia es significativa entre el X y los brazos autosómicos 2R y 3L [X vs 2R,  $P < 0,05$ ; X vs 3L,  $P < 0,01$ ; X vs 3R,  $P = 0,05$ ], y no significativa o marginalmente significativa para los brazos 2L y 3R [X vs 2L,  $P = 0,19$ ; X vs 3R,  $P = 0,09$ ], respectivamente). La fracción de mutaciones no-sinónimas ligeramente deletéreas es 1,2 veces mayor en los autosomas respecto al X (la diferencia es significativa entre el X y cada uno de los brazos autosómicos, X vs 2L,  $P < 0,05$ ; X vs 2R,  $P < 0,001$ ; X vs 3L,  $P < 0,05$ ; X vs 3R,  $P < 0,001$ ). Si comparamos ventanas autosómicas y del X con tasas de recombinación y densidad génica equivalentes encontramos que para el ADN no-codificador no hay diferencias significativas ni para la fracción de mutaciones efectivamente neutras, ligeramente deletéreas, ni fuertemente deletéreas, pero sí para la tasa de evolución adaptativa tanto para regiones intergénicas, intrones como UTRs (tabla 3.4).

## AUSENCIA DE CORRELACIÓN ENTRE LA TASA DE ADAPTACIÓN PROTEICA Y LA RECOMBINACIÓN EN EL X

Anteriormente se ha mostrado una correlación positiva no significativa entre la tasa de adaptación proteica y la recombinación en el cromosoma X (véase figura 3.7), este no es el caso para el ADN no-codificador, donde la correlación es positiva y significativa. Con tal de estudiar las posibles causas de esta ausencia de correlación se han probado distintas hipótesis.



**FIGURA 3.9** Relación entre la tasa de recombinación (cM/Mb) corregida (multiplicando las estimas de recombinación en el X por 2/3 y las de autosomas por 1/2) y la tasa de evolución adaptativa ( $\omega_A$ ) en sitios no-sinónimos estimada en ventanas no solapantes de 1 Mb – las ventanas autosómicas han sido filtradas para que en promedio tengan una densidad génica similar al conjunto de ventanas del cromosoma X. Esto se ha conseguido eliminando todas las ventanas autosómicas con una densidad génica > 14%. Las líneas son regresiones locales (LOESS).

La menor densidad génica en el X podría explicar este resultado, pues conllevaría a una reducción del efecto Hill-Robertson. La correlación entre la tasa de adaptación en sitios no-sinónimos y la tasa de recombinación para regiones autosómicas con densidades génicas similares a la del cromosoma X se muestra en la figura 3.9. Seguimos observando una correlación positiva significativa entre la tasa de

adaptación no-sinónima y la tasa de recombinación cuando regiones autosómicas de densidad génica elevada son eliminadas ( $\rho_s = 0,41, P < 0.001$ ), por lo tanto, la falta de correlación entre adaptación proteica y recombinación en el X ( $\rho_s = 0,20, P = 0,19$ ) (figura 3.7 y 3.9) no parece ser debida a la menor densidad génica en este cromosoma. Quizás el menor poder estadístico en el cromosoma X podría explicar dicha falta de correlación, nótese que en el cromosoma X disponemos de menos ventanas. Sin embargo, si elegimos al azar (mil veces sin reemplazamiento) el mismo número de ventanas autosómicas de las que disponemos en el cromosoma X seguimos observando una relación positiva significativa entre la tasa de adaptación no-sinónima y la tasa de recombinación en autosomas ( $\rho_s$  promedio = 0,41,  $P < 0.01$ ) (controlando para las diferencias en la densidad génica). Es decir, la falta de correlación en el X no parece ser debida a un menor poder estadístico en este cromosoma. Finalmente, al realizar ventanas de 1 Mb estamos reduciendo la resolución de las estimas de recombinación, las cuales han sido estimadas en ventanas de 100 kb, por lo tanto, podemos estar perdiendo más información en el cromosoma X que en los autosomas si la tasa de recombinación local cambia más rápidamente entre regiones en el X. Con tal de no perder esta información hemos agrupado los genes codificadores del cromosoma X (en 12 grupos de 136 genes cada uno) de acuerdo a la tasa de recombinación de la ventana en la que se encuentran y hemos correlacionado la tasa de adaptación con la tasa de recombinación. La correlación aumenta, pero sigue siendo no significativa ( $\rho_s = 0,26, P = 0,16$ ). En la sección 4.2.3 se da una posible explicación a esta falta de correlación entre adaptación proteica y recombinación en el cromosoma X.

### 3.3 CUANTIFICACIÓN DEL EFECTO HILL-ROBERTSON SOBRE LAS MUTACIONES

#### BENEFICIOSAS DE CAMBIO DE AMINOÁCIDO EN *Drosophila melanogaster*

A diferencia de los análisis de la sección 3.2 que estaban basados en ventanas cromosómicas no solapantes de 1 Mb y se estudiaba la huella de la selección en secuencias codificadoras y no-codificadoras, esta sección se centra sólo en secuencias codificadoras de autosomas y están basadas en agrupaciones de genes de acuerdo a la tasa de recombinación (mutación o densidad génica) de la ventana donde se encuentra dicho gen. Es decir, los genes han sido agrupados de acuerdo a una serie de propiedades no necesariamente relacionadas con su localización cromosómica. Para la mayoría de análisis se han utilizado 6.141 genes codificadores. Como en la sección anterior, las posiciones sinónimas 4-veces degeneradas han sido utilizadas como la clase de sitio neutra por excelencia, aunque para algunos análisis confirmatorios las posiciones 8 a 30 de los intrones cortos (de < 66 pb) han servido también de referencia neutra (Halligan y Keightley 2006; Parsch *et al.* 2010). Los datos de polimorfismo de una población africana de Ruanda (DPGP2, Pool *et al.* 2012) se han usado para comparar algunos de los resultados obtenidos con la población norteamericana (aunque para la población africana el tamaño de muestra es mucho menor,  $n = 17$ ).

Para estimar la tasa de evolución adaptativa se ha utilizado el método de Eyre-Walker y Keightley (2009). Como el valor de la fracción de sustituciones adaptativas ( $\alpha$ ) depende tanto de la tasa de sustitución adaptativa como no adaptativa y el estadístico  $\omega_A$  corrige para la tasa de mutación, la cual es una de las variables independientes en nuestro estudio, no hemos utilizado estos estadísticos en esta sección. En estos análisis se ha utilizado  $K_{\alpha+}$ , la tasa de sustitución aminoacídica adaptativa, la cual proviene de multiplicar la fracción de sustituciones adaptativas por la tasa de sustitución aminoacídica,  $K_{\alpha+} = \alpha K_\alpha$ .

### 3.3.1 TASA DE RECOMBINACIÓN Y ADAPTACIÓN

La figura 3.7 mostraba la relación positiva entre  $\omega_A$  y la tasa de recombinación (en cM/Mb) para los autosomas y el cromosoma X. Tanto la tasa de adaptación como la tasa de recombinación estaban estimadas en ventanas no-solapantes de 1 Mb. No obstante, la resolución de las estimas de recombinación de las que disponemos es de hasta 100 kb (Comeron *et al.* 2012), es decir, disponemos de 10 veces más resolución de la que utilizamos. En este estudio hemos querido ir más lejos y hemos agrupado los genes de acuerdo a su tasa de recombinación en 45 grupos de 136 genes cada uno para aprovechar toda la resolución del mapa de recombinación. Los resultados se muestran en la figura 3.10A. Hay una correlación positiva y significativa entre la tasa de adaptación ( $K_{a+}$ ) y la tasa de recombinación ( $\rho_s = 0.64, P < 0,001$ ). No obstante, para valores mayores a  $\sim 2$  cM/Mb, la relación entre la adaptación y la recombinación converge a un valor asintótico. Esta asintota la interpretamos como la tasa de evolución adaptativa en ausencia de interferencia de Hill-Robertson (iHR) sobre nuevas mutaciones beneficiosas. Con tal de comprobar si la relación curvilínea explica mejor los datos observados que la relación lineal hemos ajustado los datos a la función  $y = a + b \cdot e^{-cx}$  y la hemos comparado con el modelo lineal (véanse figuras S1A y S1B en el ANEXO). La tabla 3.5 muestra el valor de los parámetros, el  $R^2$  y el AIC de los dos modelos. El modelo curvilineo es estadísticamente más verosímil que el modelo lineal tal y como indica el AIC significativamente menor.

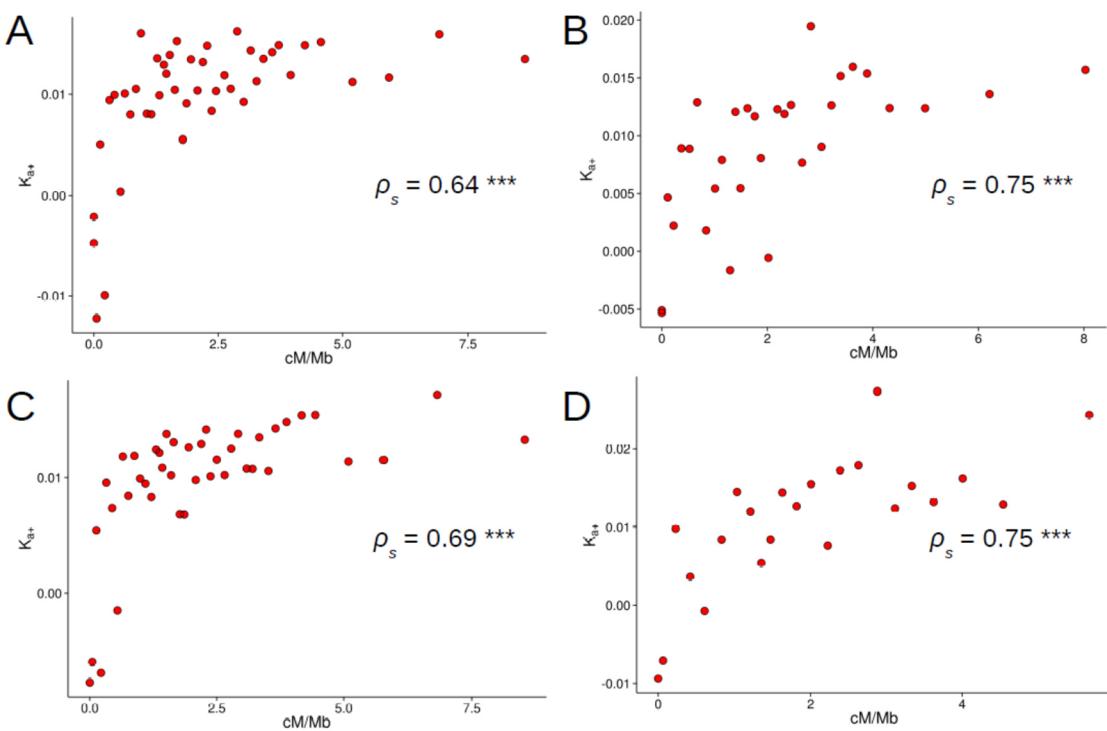
Nuestros resultados parecen *prima facie* contradecir los publicados anteriormente por Campos *et al.* (2014), donde describen una relación lineal entre  $\omega_A$  y la tasa de recombinación. Las diferencias entre ambos trabajos pueden deberse a diversas causas. Primero el estadístico utilizado para estimar la adaptación es diferente; segundo, la estrategia para agrupar los genes es distinta; tercero, los datos de polimorfismo no provienen de la misma población, y finalmente el tratamiento de las estimas de recombinación no es el mismo (véase a continuación).

**TABLA 3.5 ESTIMACIÓN DE PARÁMETROS, R<sup>2</sup> Y AIC DEL MODELO LINEAL Y CURVILINEAL PARA DISTINTOS CONJUNTOS DE DATOS DONDE Y= K<sub>x</sub> Y X= cM/Mb**

n	Linear ( $y \sim a + b \cdot x$ )			AIC	Curvilinear ( $y \sim a + b \cdot e^{cx}$ )				AIC	Pr(>F)	
	a	b	R <sup>2</sup>		a	b	c	R <sup>2</sup>			
<b>DGRP</b>	45	0.0058	0.0018	27%	-336.85	0.0127	-0.0186	2.1237	67%	-369.78	1.36E-8
<b>DGRP</b>	11	0.0058	0.0019	52%	-88.58	0.0132	-0.0144	1.4382	94%	-110.07	5.62E-5
<b>DPGP2</b>	31	0.0042	0.0021	38%	-236.78	0.0149	-0.0148	0.5843	51%	-241.75	1.29E-2
<b>DPGP2</b>	11	0.0036	0.0024	54%	-84.45	0.0144	-0.0156	0.7429	72%	-88.01	5.11E-2
w/o IT	42	0.0059	0.0017	32%	-327.03	0.0124	-0.0183	2.0991	69%	-358.04	3.72E-8
<b>High Fop</b>	15	0.0026	0.0017	52%	-125.34	0.0097	-0.0145	1.2983	91%	-148.41	1.09E-5
<b>Med Fop</b>	15	0.0081	0.0009	63%	-150.09	0.0135	-0.0069	0.4014	71%	-151.74	9.40E-2
<b>Low Fop</b>	15	0.0037	0.0033	43%	-101.74	0.0162	-0.0270	1.7242	80%	-115.49	4.94E-4
<b>Short Int</b>	23	0.0026	0.0041	51%	-164.96	0.0175	-0.0229	0.9801	67%	-171.86	5.98E-3
<b>GenH-MutH</b>	12	0.0009	0.0034	56%	-86.69	0.0193	-0.0222	0.4158	62%	-86.55	2.50E-1
<b>GenH-MutL</b>	12	0.0030	0.0016	26%	-91.75	0.0078	-0.1039	18.6800	90%	-114.09	2.98E-5
<b>GenL-MutH</b>	12	0.0076	0.0031	37%	-76.76	0.0219	-0.0280	1.2229	84%	-91.26	5.98E-4
<b>GenL-MutL</b>	12	0.0054	0.0009	27%	-99.33	0.0095	-0.0128	2.0985	87%	-118.20	1.12E-4

La primera columna (n) corresponde al número de agrupaciones de genes. El valor p de la prueba F utilizada para comparar los dos modelos se encuentra en la última columna. El conjunto de datos DGRP original (población norteamericana) se encuentra en las filas 1-2. El conjunto de datos DPGP2 original (población africana) está en las filas 3-4. En la fila 5 (w/o IT) se han excluido los genes del sistema inmune y de expresión sesgada en machos. Las filas 6-8 muestran los resultados para los genes con un elevado (High Fop), intermedio (Med Fop) y bajo (Low Fop) Fop, respectivamente. La fila 9 muestra los resultados utilizando los intrones cortos como referencia neutra. Las filas 10-13 corresponden a los genes GenH-MutH, GenH-MutL, GenL-MutH, y GenL-MutL, respectivamente, consultese sección 3.3.4 para la definición de estos grupos.

Las diferencias entre ambos estudios no son atribuibles al estadístico utilizado para estimar la adaptación (véanse figuras 3.10A y S2A en el ANEXO), pues tanto para  $K_{\alpha}$ , como para  $\omega_A$  el modelo curvilíneo es más verosímil (véase tabla 3.5). La estrategia utilizada para agrupar los genes tampoco parece la razón detrás de las diferencias, pues cuando agrupamos los genes del mismo modo que Campos *et al.* (2014) (10 grupos con tasas de recombinación  $> 0$  cM/Mb y un grupo con  $0$  cM/Mb) observamos que la relación curvilínea es significativamente mejor que la lineal (véase figura S3 A-B en el ANEXO y tabla 3.5). Campos *et al.* (2014) utilizó dos estimas de la tasa de recombinación: una basada en marcadores cromosómicos visibles de baja resolución (Fiston-Lavier *et al.* 2010) y la otra utiliza el mapa de alta resolución basado en SNPs como marcadores (Comeron *et al.* 2012). En ambos casos, Campos *et al.* (2014) efectuó un ajuste lineal de los datos. No obstante, en lugar de utilizar directamente las estimas puntuales de la tasa de recombinación descrita por Comeron *et al.* (2012) en ventanas no solapantes de 100 kb, como hemos hecho nosotros, Campos *et al.* (2014) hizo una regresión local (LOESS) para suavizar los patrones de recombinación a lo largo de los cromosomas. Esta regresión disminuye la resolución del mapa original y podría estar detrás de las diferencias entre ambos trabajos. Hemos repetido la correlación de Campos *et al.* (2014) con sus datos genómicos (esto es, el polimorfismo proveniente de una población africana [DPGP2, Pool *et al.* 2012] y la divergencia con *D. yakuba*) y el mapa de recombinación de alta resolución original de Comeron *et al.* (2012). En esta ocasión recuperamos la relación curvilínea que muestra la población norteamericana (figura 3.10B, tabla 3.5 y figura S3 C-D del ANEXO) ( $\rho_s = 0,75$ ,  $P < 0,001$ ). En definitiva, la regresión local del mapa de recombinación parece generar una relación lineal artefactual entre la tasa de adaptación y la tasa de recombinación. Ni el estadístico de la adaptación, ni la estrategia seguida para agrupar los genes, ni la población utilizada para estimar el polimorfismo son por tanto la explicación.



**FIGURA 3.10** Relación entre  $K_{\alpha^+}$  en el eje de la y y la tasa de recombinación (cM/Mb) en el eje de la x: (A) utilizando el polimorfismo de una población norteamericana (DGRP), (B) utilizando el polimorfismo de una población africana (DPGP2), (C) excluyendo los genes del sistema inmune y de expresión sesgada en machos y (D) utilizando los intrones cortos como clase de sitios neutros. Cada punto se ha estimado agrupando genes. El número de genes, la recombinación promedio y la estima de  $K_{\alpha^+}$  para cada grupo puede consultarse en la tabla S1 del ANEXO.  $\rho_s$ : coeficiente de la correlación de Spearman, la significación se denota con asteriscos (\*\*<0.001; \*\*<0.01; \*<0.05).

Por todas estas razones a partir de ahora sólo se mostrarán los resultados obtenidos con la población norteamericana (Mackay *et al.* 2012), además de por la mayor cobertura y tamaño de muestra de esta población respecto a la población africana (Pool *et al.* 2012).

La relación positiva entre la tasa de recombinación y  $K_{\alpha^+}$  podría deberse a otras causas no directamente relacionadas con la iHR. Si la recombinación es mutagénica esperaríamos una relación positiva entre la tasa de recombinación y  $K_{\alpha^+}$ . No obstante, estudios previos no han encontrado evidencias a favor de esta hipótesis (Begun y Aquadro 1992; Begun *et al.* 2007; McGaugh *et al.* 2012) y nosotros mismos no

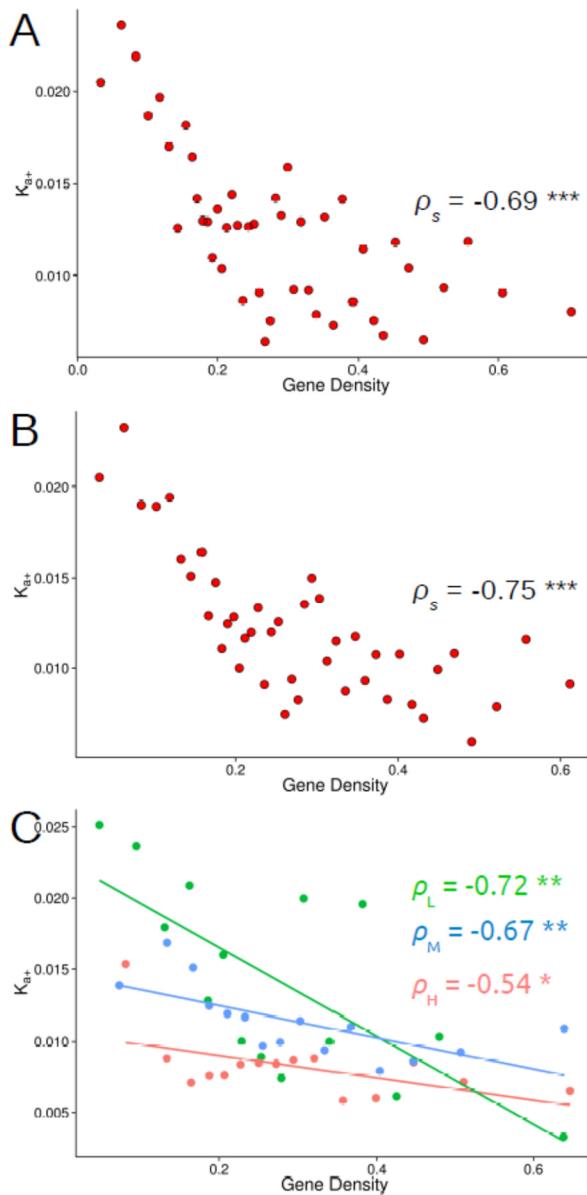
encontramos correlación entre la tasa de substitución en intrones cortos, los cuales se asume son el tipo de sitios más neutros en *Drosophila* (Halligan y Keightley 2006; Parsch *et al.* 2010) y la tasa de recombinación ( $\rho_s = 0,01, P = 0,89$ ). El enriquecimiento de genes con elevadas tasas de adaptación, como los genes del sistema inmune (Obbard *et al.* 2009) y de expresión sesgada en machos (Pröschel *et al.* 2006; Haerty *et al.* 2007), en regiones del genoma con elevada recombinación generaría una relación positiva entre la adaptación y la recombinación no necesariamente debido a la iHR. Hemos querido confirmar si este tipo de genes tienen por un lado una mayor tasa de evolución adaptativa y por otro lado si estos están afectando nuestra correlación entre  $K_{\alpha}$  y la tasa de recombinación. Estudios previos no corregían para la presencia de alelos ligeramente deletéreos en las estimas de adaptación, mientras que el estimador utilizado por nosotros sí. Hemos encontrado que los genes del sistema inmune y de expresión sesgada en machos muestran tasas de evolución adaptativa 1,37 veces mayores que el resto de genes (véase figura S4 en el ANEXO) (Prueba de permutación  $P < 0,05$ ). Si eliminamos estos genes de nuestro conjunto de datos todavía observamos una relación curvilínea y significativa entre la tasa de recombinación y  $K_{\alpha}$  (figura 3.10C y tabla 3.5) ( $\rho_s = 0,69, P < 0,001$ ).

Al estimar la tasa de evolución adaptativa hemos asumido que las mutaciones sinónimas son neutras, sin embargo, es conocido que la selección actúa sobre las mutaciones sinónimas en *Drosophila* (revisado por Hershberg y Petrov 2008). Generalmente se ha pensado que la selección favorece aquellos codones que se traducen más rápidamente o con una menor tasa de error (Shields *et al.* 1988; Akashi 1994, 1995; Carlini y Stephan 2003; Stoletzki y Eyre-Walker 2006). Aunque las posiciones sinónimas pueden estar bajo selección para mantener o evitar intensificadores del corte y empalme (*splicing enhancers*) (Parmley *et al.* 2006), estructuras secundarias en el RNA (Parsch *et al.* 1997; Baines *et al.* 2004; Stoletzki 2008) o motivos cortos de reconocimiento para proteínas (Antezana y Kreitman 1999). Lawrie *et al.* (2013) mostraron que un ~22% de todas las posiciones sinónimas 4-

veces degeneradas en *D. melanogaster* están bajo selección purificadora fuerte, aunque el mecanismo funcional detrás de este fuerte constreñimiento se desconoce. En este análisis también confirmamos resultados previos en los que  $K_{a+}$  se encontraba significativa y negativamente correlacionada con una medida del sesgo en el uso de codones, *Fop* (*the frequency of optimal codons*) ( $\rho_s = -0,4, P < 0,001$ ) (Sharp y Li 1987, 1989; Moriyama y Hartl 1993; Bierne y Eyre-Walker 2003; 2006). En cualquier caso, esperamos que cualquier tipo de selección purificadora débil que pueda actuar sobre nuevas mutaciones sinónimas genere una correlación positiva entre  $K_{a+}$  y la tasa de recombinación. Esto es así porque los genes localizados en regiones de alta recombinación, donde la selección negativa es más eficiente (Kliman y Hey 1993; Haddrill *et al.* 2007; Campos *et al.* 2012), tenderán a tener estimas de  $K_{a+}$  elevadas porque la selección negativa débil reduce más la divergencia sinónica que el polimorfismo sinónico. Por lo tanto, para investigar si la selección en el uso de codones está afectando nuestros resultados hemos dividido nuestro conjunto de datos en 3 grupos de acuerdo a sus niveles de *Fop*, y dentro de cada uno de estos grupos hemos hecho 15 subgrupos de genes de acuerdo a sus niveles de recombinación. En total, se han analizado 45 grupos de 136 genes cada uno (como en el análisis anterior). Para cada uno de las tres categorías de *Fop* observamos una relación curvilínea significativa (véase tabla 3.5 y tabla S5 en el ANEXO) y una fuerte correlación positiva (para genes con *Fop* alto  $\rho_s = 0,87, P < 0,001$ ; *Fop* intermedio  $\rho_s = 0,78, P < 0,001$ ; *Fop* bajo  $\rho_s = 0,76, P < 0,001$ ) entre  $K_{a+}$  y la tasa de recombinación. Hemos utilizado un conjunto de datos menor con 3.369 genes donde el polimorfismo y las substituciones en intrones cortos (< 66 pb) han sido utilizados como referencia neutra. Este conjunto de datos es menor porque no todos los intrones cumplen los criterios de tamaño y calidad impuestos. Observamos la misma relación curvilínea (véanse figura 3.10D y tabla 3.5) y la intensidad de la correlación es equivalente a la encontrada con los sitios 4-veces degenerados ( $\rho_s = 0,75, P < 0,001$ ). Por lo tanto, la selección en el uso de codones no parece estar detrás ni de la forma ni de la intensidad de la relación entre la tasa de adaptación y la tasa de recombinación.

### 3.3.2 DENSIDAD GÉNICA Y ADAPTACIÓN

La intensidad de la iHR se espera que dependa tanto de la tasa de recombinación como de la densidad de sitios bajo selección a lo largo del genoma. Esperamos por lo tanto una relación negativa entre la tasa de adaptación y la densidad génica, una relación que observamos (figura 3.11A) ( $\rho_s = -0,69, P < 0,001$ ). La relación permanece después de excluir los genes del sistema inmune y de expresión sesgada en machos (figura 3.11B) ( $\rho_s = -0,75, P < 0,001$ ). No obstante, contrario a lo que esperamos, encontramos, al igual que Hey y Kliman (2002), una relación positiva débil entre el sesgo en el uso de codones y la densidad génica ( $\rho_s = 0,07, P < 0,001$ ). Esto es curioso porque regiones de alta densidad génica se espera tengan más iHR y como consecuencia una menor eficiencia de la selección en el uso de codones. Hey y Kliman (2002) encontraron que los niveles de sesgo en el uso de codones y de expresión son máximos para aquellos genes situados en un intervalo intermedio de densidad génica, un patrón que puede resultar del compromiso entre la ventaja que supone expresar a niveles elevados genes compactos y la desventaja que surge al tener que regular genes con escasa secuencia no-codificadora entre ellos. En cualquier caso, para comprobar que esta correlación positiva entre densidad génica y *Fop* no está induciendo una correlación negativa artefactual entre  $K_{a+}$  y la densidad génica hemos dividido nuestros genes en tres categorías de acuerdo a su *Fop*. Para cada uno de los tres niveles de *Fop* encontramos una relación negativa significativa (figura 3.11C) (para genes con *Fop* alto  $\rho_s = -0,54, P < 0,05$ ; *Fop* intermedio  $\rho_s = -0,67, P < 0,01$ ; *Fop* bajo  $\rho_s = -0,72, P < 0,01$ ). Obtenemos resultados equivalentes entre  $\omega_A$  y la densidad génica (véase figura S6 en el ANEXO).

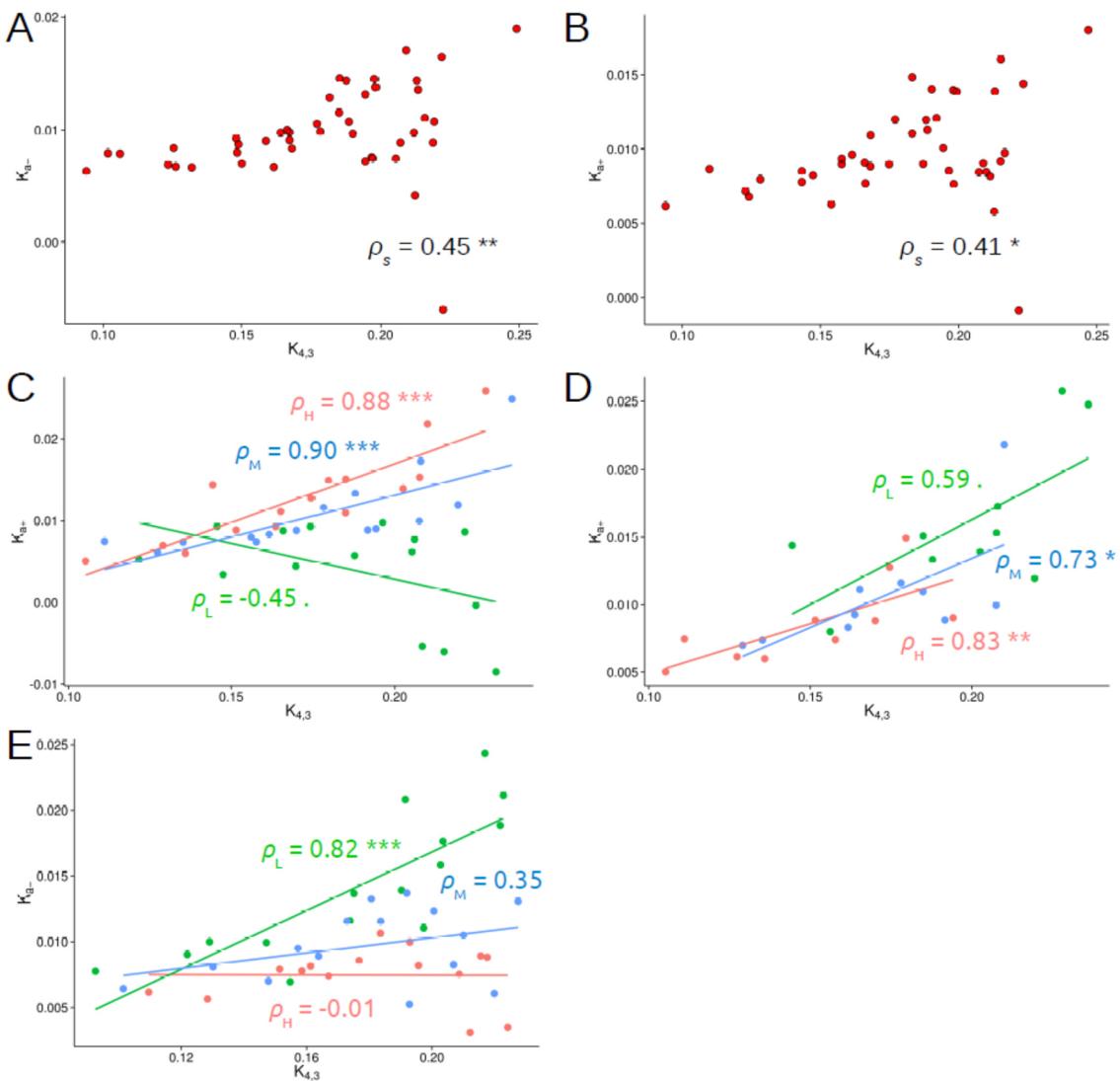


**FIGURA 3.11** Relación entre  $K_{\alpha^+}$  en el eje de la y y la densidad génica en el eje de la x: (A) utilizando el conjunto de datos completo, (B) excluyendo genes del sistema inmune y de expresión sesgada en machos y (C) diferenciando entre niveles de *Fop*. La relación para los genes que pertenecen a la categoría de elevado *Fop* (H) están coloreados en rojo, con *Fop* intermedio (M) en azul y con bajo *Fop* (L) en verde. Cada punto se ha estimado agrupando genes. El número de genes, la densidad génica y *Fop* promedio y la estima de  $K_{\alpha^+}$  para cada grupo puede consultarse en la tabla S4 del ANEXO.  $\rho_s$ : coeficiente de la correlación de Spearman, la significación se denota con asteriscos (\*\*<0.001; \*\*<0.01; \*<0.05). Las líneas corresponden a regresiones por mínimos cuadrados.

### 3.3.3 TASA DE MUTACIÓN Y ADAPTACIÓN

Estudiar la relación entre la tasa de mutación y la tasa de adaptación no es tan sencillo como hacer una correlación directa entre estas dos variables. La tasa de substituciones sinónimas se utiliza para estimar tanto la tasa de mutación como la tasa de adaptación, es decir, estas dos variables no son independientes. Esta falta de independencia estadística entre estimas genera una correlación negativa entre  $K_{\alpha+}$  y  $K_4$  simplemente debido al error de muestreo. Para evitar esta complicación, se ha muestreado tres veces consecutivas y sin reemplazamiento las substituciones sinónimas a nivel de gen (para conocer los detalles de este re-muestreo consultese sección 2.6). Esto ha permitido obtener tres estimas independientes de  $K_4$  a nivel de gen:  $K_{4,1}$  ha sido utilizada para ordenar los genes y asignarlos a los distintos grupos,  $K_{4,2}$  se ha usado para estimar la tasa de evolución adaptativa, y  $K_{4,3}$  es la estima de la tasa de mutación para cada grupo de genes.

Como en los análisis anteriores, el conjunto de datos fue dividido en 45 grupos de 136 genes cada uno pero después de ordenar los genes por su valor de  $K_{4,1}$ . Haciendo esto encontramos una correlación positiva entre  $K_{\alpha+}$  y  $K_{4,3}$  ( $\rho_s = 0,45, P < 0,001$ ) (figura 3.12A). Esta correlación se mantiene después de eliminar los genes relacionados con el sistema inmune o de expresión sesgada en machos ( $\rho_s = 0,41, P < 0,01$ ) (figura 3.12B), lo cual sugiere que la correlación entre  $K_{\alpha+}$  y  $K_{4,3}$  no es una consecuencia de que estos genes muten más. La selección en el uso de codones se espera genere una correlación negativa entre  $K_{\alpha+}$  y  $K_{4,3}$  porque la selección negativa débil reduce más la divergencia que el polimorfismo. Con tal de investigar como la selección en los sitios sinónimos afecta a la correlación entre  $K_{\alpha+}$  y  $K_{4,3}$  hemos dividido los genes en 3 grandes grupos de acuerdo a su tasa de recombinación y su valor de *Fop* (en total tenemos 9 grupos). Cada uno de estos 9 grupos se han ordenado y dividido en 5 de acuerdo a  $K_{4,1}$  (en total tenemos de nuevo 45 grupos de 136 genes cada uno).



**FIGURA 3.12** Relación entre  $K_{\theta_r}$  en el eje de la y y una estimación de la tasa de mutación ( $K_{4,3}$ ) en el eje de la x: (A) utilizando todo el conjunto de datos, (B) excluyendo genes del sistema inmune y de expresión sesgada en machos, (C) diferenciando entre niveles de recombinación, (D) diferenciando los genes por su sesgo en el uso de codones (*Fop*) después de eliminar los genes de baja recombinación (< 1.32 cM/Mb) y (E) diferenciando entre niveles de densidad génica. Los genes pertenecientes a las categorías altas (H) están coloreadas en rojo, los genes de categorías intermedias (M) están coloreadas en azul y los genes de categorías bajas (L) están en verde. Cada punto se ha estimado agrupando genes. El número de genes y el valor de los distintos estadísticos para cada grupo puede consultarse en la tabla S5 del ANEXO.  $\rho_s$ : coeficiente de la correlación de Spearman, la significación se denota con asteriscos (\*\*<0.001, \*\*<0.01; \*<0.05). Las líneas corresponden a regresiones por mínimos cuadrados.

Hemos decidido tener en cuenta la tasa de recombinación porque esta afecta tanto a la tasa de adaptación como a la eficiencia de la selección en el uso de codones. La correlación entre  $K_{a+}$  y  $K_{4,3}$  para cada categoría de recombinación y de *Fop* se encuentra en la figura 3.12 C y D, respectivamente. Los gráficos sugieren que la selección en el uso de codón afecta poco a la correlación entre  $K_{a+}$  y  $K_{4,3}$ , sin embargo, esta relación se ve muy afectada por la tasa de recombinación. Este hecho no es sorprendente, anteriormente se ha mostrado que genes situados en regiones de muy baja recombinación muestran escasas evidencias de adaptación (véase figura 3.10), por lo tanto, es esperable que estos genes estén poco influenciados por su tasa de mutación.

Para investigar estas cuestiones más formalmente hemos realizado un análisis de la covarianza (ANCOVA), agrupando a los genes por sus niveles de *Fop* y tasa de recombinación. En un ANCOVA se ajustan un conjunto de rectas sobre los datos, una por grupo. Esto permite comprobar estadísticamente si la pendiente y el intercepto de estas rectas son diferentes entre ellas. Si consideramos a *Fop* y la tasa de recombinación como factores fijos no encontramos evidencias de que  $K_{a+}$  y  $K_{4,3}$  estén correlacionados (ANCOVA  $P = 0,16$ ). Sin embargo, encontramos que las pendientes (ANCOVA  $P < 0,001$ ) y los interceptos (ANCOVA  $P < 0,001$ ) difieren entre categorías de recombinación. Estas diferencias significativas no se observan para las distintas categorías de *Fop*. Si los genes pertenecientes a la categoría de baja recombinación (de 0 a 1,32 cM/Mb) son excluidos, el resto de genes muestran una fuerte correlación positiva entre  $K_{a+}$  y  $K_{4,3}$  (véase figura S7 en el ANEXO) ( $\rho_s = 0,82$ ,  $P < 0,001$ ). Para este conjunto de datos no hay evidencias de que la pendiente o intercepto difiera de acuerdo a la tasa de recombinación o el sesgo en el uso de codón. Como aproximación alternativa para controlar el efecto de la selección en el uso de codón en nuestras estimas de la tasa de mutación, hemos hecho una regresión donde  $K_{4,3}$  es la variable dependiente y la tasa de recombinación y *Fop* son las variables independientes. Con los residuos de dicha regresión encontramos una fuerte correlación positiva con  $K_{a+}$ .

(véase figura S8 en el ANEXO) ( $\rho_s = 0,42, P < 0,01$ ), hecho que sugiere, de nuevo, que la correlación entre  $K_{a+}$  y  $K_{4,3}$  no es el resultado de la selección débil en las posiciones 4-veces degeneradas.

Se ha mostrado como la tasa de recombinación afecta la relación entre  $K_{a+}$  y  $K_{4,3}$ , y es lógico esperar que la densidad génica tenga un efecto similar. La relación entre  $K_{a+}$  y  $K_{4,3}$  será más fuerte en regiones del genoma con menor densidad génica. Para investigar esta cuestión, el conjunto de datos se ha dividido en tres niveles de densidad génica y dentro de cada grupo los genes se han ordenado y dividido en 15 subgrupos de acuerdo a su tasa de mutación (45 grupos de 136 genes cada uno han sido analizados). La figura 3.12E muestra la relación entre la tasa de mutación y la tasa de adaptación para cada nivel de densidad génica. Si consideramos a la densidad génica como un factor fijo encontramos mediante un ANCOVA una correlación positiva y significativa entre  $K_{a+}$  y  $K_{4,3}$  (ANCOVA  $P < 0,01$ ). No obstante, encontramos diferencias significativas entre las pendientes (ANCOVA  $P < 0,01$ ) y los interceptos (ANCOVA  $P < 0,05$ ) entre niveles de densidad génica. Cuando los genes situados en regiones de baja densidad génica son excluidos no encontramos correlación entre  $K_{a+}$  y  $K_{4,3}$  (ANCOVA  $P = 0,51$ ) ni evidencias de diferencias en las pendientes (ANCOVA  $P = 0,70$ ) ni los interceptos (ANCOVA  $P = 0,13$ ) entre los grupos de densidad génica media y alta. En definitiva, sólo observamos una correlación positiva significativa entre  $K_{a+}$  y  $K_{4,3}$  para los genes de baja densidad génica ( $\rho_s = 0,82, P < 0,001$ ) y una correlación positiva no significativa para el resto de genes (figura 3.12E).

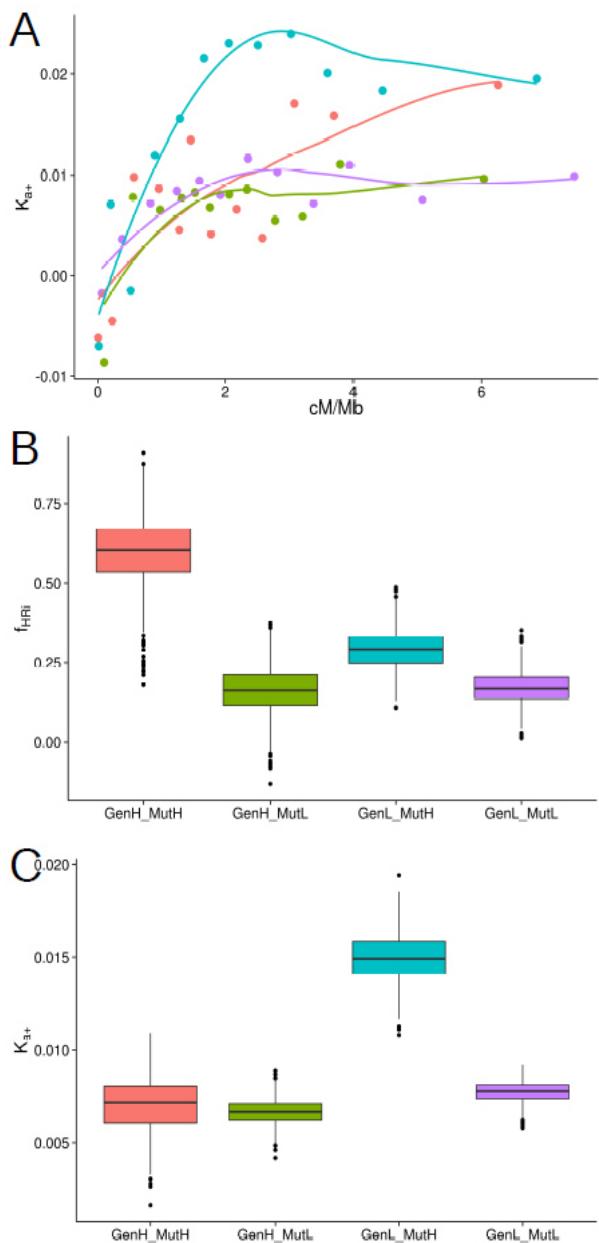
Nuestros resultados muestran como globalmente existe una correlación significativa y positiva entre la tasa de mutación y la tasa de adaptación (véanse figura 3.12 A-B y figuras S7 y S8 en el ANEXO), porque los genes con elevadas tasas de mutación es más probable que generen la variación genética necesaria para la adaptación. Esto no quiere decir que la relación positiva pueda aplicarse a cualquier contexto genómico. De hecho, la fuerza y el signo de la relación entre  $K_{a+}$  y  $K_{4,3}$  depende de la

tasa de recombinación (figura 3.12C) y de la densidad génica (figura 3.12E). Hemos mostrado como cuando la densidad génica es alta y/o la tasa de recombinación es baja la correlación entre la tasa de mutación y la tasa de adaptación es muy baja o inexistente debido a la iHR.

### 3.3.4 LA PROPORCIÓN DE MUTACIONES ADAPTATIVAS NO FIJADAS DEBIDO AL EFECTO HILL-ROBERTSON

Nuestros resultados muestran que la tasa de adaptación es significativamente menor en el genoma de *Drosophila* en regiones de baja recombinación y elevada densidad génica. Pero, ¿cuántas substituciones adaptativas dejan de fijarse debido a la iHR? Para una tasa de recombinación dada, ¿cómo la tasa de mutación y la densidad génica afectan la intensidad de la iHR? Para responder a estas preguntas hemos realizado una regresión local (LOESS) entre  $K_{\alpha+}$  y la tasa de recombinación. Esta relación converge a una asíntota por encima de 2 cM/Mb (véase figura S1C en el ANEXO). Interpretamos el valor de la asíntota (2cM/Mb) como la tasa de evolución adaptativa que ocurre en ausencia de iHR. La diferencia entre el valor de  $K_{\alpha+}$  asintótico y el encontrado < 2 cM/Mb lo interpretamos como el número de sustituciones que se pierden debido a la iHR. Utilizando esta metodología, y ponderando por el número de sitios involucrados, hemos estimado que un 27,2% (intervalos de confianza al 95% obtenidos mediante *bootstrap* [20,6%, 33,8%]) de todas las sustituciones beneficiosas de cambio de aminoácido que deberían fijarse en un genoma con libre recombinación se han perdido debido a la iHR. Designamos a esta proporción de sustituciones adaptativas perdidas debido a la iHR como  $f_{HRI}$ . Algunas de las estimas de  $K_{\alpha+}$  de la regresión local son negativas (sobre todo las de baja recombinación), sin embargo, nuestra estima de la  $f_{HRI}$  no parece verse afectada si cambiamos estos valores negativos a ceros,  $f_{HRI} = 27,1\%$  (95% IC [20,6%, 33,2%]).

No obstante, la iHR se espera que sea más prevalente en aquellos *loci* con altas tasas de mutación y/o en aquellos *loci* localizados en regiones de alta densidad génica, porque esto aumentará la probabilidad de que una mutación seleccionada se encuentre segregando con otras mutaciones sometidas a selección. Para investigar si estas predicciones se cumplen en el genoma de *Drosophila* hemos repetido el análisis anterior pero separando los genes en distintas categorías de acuerdo a su tasa de mutación y la densidad génica de la ventana cromosómica donde están localizados. Primero, hemos dividido nuestro conjunto de datos en dos partes iguales de acuerdo a la densidad génica, y dentro de cada uno de estos dos grupos los genes se han separado de acuerdo a su tasa de mutación. Resultados cualitativamente equivalentes son obtenidos si primero sepáramos por la tasa de mutación y luego por la densidad génica. Para separar los genes según su tasa de mutación primero debemos generar dos estimas estadísticamente independientes de  $K_{a+}$ , la primera ( $K_{4,1}$ ) servirá para ordenar y separar a los genes por su tasa de mutación, y la segunda ( $K_{4,2}$ ) servirá para estimar la tasa de adaptación. Esto lo hacemos, de nuevo, muestreando las substituciones sinónimas a partir de una distribución hipergeométrica (consúltese sección 2.6 para más detalles). Haciendo esto nos aseguramos que nuestras estimas de  $K_{a+}$  no se ven influenciadas por la manera en la que los datos son divididos. Hemos etiquetado estos 4 grupos de genes de la siguiente manera (por sus siglas en inglés): GenH-MutH (*high gene density and high mutation rate genes*), GenH-MutL (*high gene density and low mutation rate genes*), GenL-MutH (*low gene density and high mutation rate genes*) y GenL-MutL (*low gene density and low mutation rate genes*). La relación entre  $K_{a+}$  y la tasa de recombinación para cada una de estas categorías se puede encontrar en la figura 3.13 A. La fuerza de la relación es equivalente a la encontrada previamente con el conjunto completo de datos (GenH-MutH genes  $\rho s = 0,67, P < 0,05$ ; GenH-MutL genes  $\rho s = 0,48, P < 0,05$ ; GenL-MutH genes  $\rho s = 0,67, P < 0,05$  y GenL-MutL genes  $\rho s = 0,55, P < 0,05$ ). No obstante, la relación para los genes GenH-MutH parece lineal, mientras que para el resto de categorías la relación curvilínea es significativa (tabla 3.5).



**FIGURA 3.13** (A) Relación entre  $K_{\alpha^+}$  en el eje de la y y la tasa de recombinación (cM/Mb) en el eje de la x para cada categoría de genes; las líneas son regresiones locales. (B) Valores de  $f_{HRI}$  y (C) de  $K_{\beta^+}$  obtenidos mediante *bootstrap* para cada categoría de genes. Cada punto ha sido estimado agrupando 128 genes de acuerdo a su densidad génica, tasa de mutación ( $K_{\beta^+}$ ) y tasa de recombinación (cM/Mb). Los genes GenH-MutH están coloreados en rojo, los genes GenH-MutL en verde, los genes GenL-MutH en azul y los genes GenL-MutL en violeta; consultese texto principal para una descripción completa de estas categorías.

Como la relación entre  $K_{\alpha+}$  y la tasa de recombinación para los genes GenH-MutH no muestra una aproximación asintótica, sólo podemos intentar estimar la proporción mínima de substituciones perdidas debido a la iHR. Para hacer esto asumimos que los genes GenH-MutH por encima de 5 cM/Mb no sufren iHR, para el resto de categorías seguimos utilizando el valor umbral de 2 cM/Mb. Encontramos que la  $f_{HRI}$  varía significativamente entre categorías de genes: genes con elevadas tasas de mutación localizados en regiones ricas en genes pierden más substituciones beneficiosas que genes pertenecientes a las otras categorías (GenH-MutH vs: GenH-MutL *bootstrap P < 0,01*, GenL-MutH *bootstrap P < 0,05*, GenL-MutL *bootstrap P < 0,01*), los cuales no son significativamente diferentes entre ellos. Hemos estimado que los genes GenH-MutH han perdido ~59,7% (95% ICs [41,5%, 75,6%]) de todas sus substituciones debido a la iHR comparado con un 20% aproximadamente en las otras categorías (figura 3.13B y tabla 3.6). Si calculamos la fracción global de substituciones perdidas debido a la iHR combinando los datos de las 4 categorías encontramos que ~35,9% (95% ICs [27,0%, 44,2%]) de todas las substituciones de cambio de aminoácido que se deberían de haber fijado en un genoma donde los sitios son independientes se han perdido debido a la iHR. En cualquier caso, esta nueva estima genómica de la  $f_{HRI}$  no es significativamente diferente a la estima anterior la cual rondaba el ~27% (*bootstrap P = 0,18*).

Finalmente, aunque hay variación en la fracción de sustituciones adaptativas perdidas debido a la iHR entre categorías de genes, ¿cuántas sustituciones beneficiosas acaban fijándose? La figura 3.13C muestra la tasa de evolución adaptativa para cada categoría de genes en cada una de las 1000 réplicas del *bootstrap*. Los genes que muestran una mayor tasa de evolución adaptativa son aquellos con elevadas tasas de mutación localizados en regiones de baja densidad génica (genes GenL-MutH)  $K_{\alpha+} = 0,0149$  (95% ICs [0,0128, 0,0171]), mientras que el resto de categorías muestran niveles similares de adaptación: GenH-MutH  $K_{\alpha+} = 0,007$  (95% ICs [0,0044, 0,0092]), GenH-MutL  $K_{\alpha+} = 0,0067$  (95% ICs [0,0054, 0,0069]) y GenL-MutL  $K_{\alpha+} = 0,0077$  (95% ICs [0,0067, 0,0086]).

**TABLA 3.6** MEDIA E IC 95% OBTENIDOS MEDIANTE *BOOTSTRAP* PARA DIVERSOS ESTADÍSTICOS

Category	K <sub>a+</sub>	K <sub>4,2</sub>	f <sub>HRI</sub>	AA+ (kb)	Lost AA+ (kb)	Global f <sub>HRI</sub>
GenH-MutH	0.0070 (0.0045, 0.0092)	0.24 (0.23, 0.25)	0.60 (0.41, 0.76)	9.97 (6.37, 13.07)	15.46 (7.74, 22.77)	
GenH-MutL	0.0067 (0.0054, 0.0079)	0.18 (0.17, 0.18)	0.16 (0.02, 0.28)	9.58 (7.83, 11.34)	1.87 (0.26, 3.37)	
GenL-MutH	0.0149 (0.0128, 0.0171)	0.24 (0.23, 0.24)	0.29 (0.19, 0.40)	18.93 (16.17, 21.69)	7.83 (4.80, 11.17)	0.36 (0.27, 0.44)
GenL-MutL	0.0077 (0.0067, 0.0086)	0.17 (0.16, 0.17)	0.17 (0.08, 0.27)	9.50 (8.29, 10.62)	1.97 (0.88, 3.19)	

K<sub>a+</sub>, tasa de substituciones no sinónimas adaptativas (corregida por el método Jukes y Cantor 1969); K<sub>4,2</sub>, tasa de substitución en posiciones 4-veces degeneradas (corregida por el método de Tamura 1992); f<sub>HRI</sub>, fracción de substituciones adaptativas perdidas debido a la iHR; AA+ (kb), número absoluto de fijaciones adaptativas de cambio de aminoácido fijadas (en kb); lost AA+ (kb), número absoluto de fijaciones adaptativas de cambio de aminoácido que se han perdido debido a la iHR (en kb) y Global f<sub>HRI</sub>, fracción global (combinando los 4 grupos) de substituciones adaptativas perdidas debido a la iHR.

Por lo tanto, aunque los genes GenH-MutH y GenL-MutH tienen tasas de mutación significativamente mayores a los genes GenH-MutL y GenL-MutL (K<sub>4,2</sub> fold-change = 1,4, bootstrap P < 0,001), los genes GenH-MutH pierden muchas más mutaciones adaptativas que los genes con bajas tasas de mutación (GenH-MutL y GenL-MutL), y como consecuencia la tasa de evolución adaptativa acaba siendo prácticamente la misma para los genes GenH-MutH, GenH-MutL y GenL-MutL (GenH-MutH vs GenH-MutL bootstrap P = 0,41; GenH-MutH vs GenL-MutL bootstrap P = 0,35; GenH-MutL vs GenL-MutL bootstrap P = 0,13). En cambio, los genes GenL-MutH son menos susceptibles a la iHR debido a la baja densidad génica y en definitiva pueden adaptarse más rápidamente que el resto de genes debido a sus elevadas tasas de mutación (GenL-MutH versus: GenH-MutH bootstrap P < 0,001, GenH-MutL bootstrap P < 0,001, GenL-MutL bootstrap P < 0,001).

Esta tesis consta de tres partes diferenciadas, la primera corresponde al desarrollo y aplicación sobre el genoma de *D. melanogaster* de dos nuevos estimadores de la acción de la selección purificadora (sección 4.1). La bondad de los estimadores se ha probado con datos simulados generados a partir de ecuaciones de difusión y simuladores *forward in time* (Hernandez 2008). En la segunda parte se dan las estimas más precisas y completas actualmente de la *DFE* de nuevas mutaciones deletéreas y la tasa de evolución adaptativa en el genoma codificador y no-codificador de *D. melanogaster* (véase sección 4.2). En la tercera y última parte se ha llevado a cabo un análisis estadístico a nivel genómico donde se ha estudiado el papel de los determinantes genéticos del efecto Hill-Robertson (estos son la tasa de recombinación, la densidad génica y la tasa de mutación) sobre la tasa de substituciones adaptativas de cambio de aminoácido en *D. melanogaster* (véase sección 4.3). Además, se ha estimado, por primera vez en una especie, qué proporción de las substituciones adaptativas de cambio de aminoácido dejan de fijarse como consecuencia de la interferencia de Hill-Robertson.

Es importante destacar que la disponibilidad de 158 genomas de alta calidad, con una alta cobertura, de una población natural de la especie *D. melanogaster* (Mackay *et al.* 2012) nos ha permitido realizar uno de los estudios más exhaustivos hasta la fecha en cuanto a la detección y cuantificación de la huella de la selección natural y el efecto Hill-Robertson a lo largo de un genoma. A su vez, la existencia de potentes métodos de inferencia de la *DFE* y la tasa de evolución adaptativa (Keightley y Eyre-Walker 2007; Eyre-Walker y Keightley 2009) y la disponibilidad del mapa de recombinación actual de alta resolución (Comeron *et al.* 2012) han sido otros recursos indispensables.

## 4.1 ESTIMADORES $d_n$ Y $b$ : NUEVAS PRUEBAS ESTADÍSTICAS DE LA ACCIÓN DE LA SELECCIÓN PURIFICADORA

Como se adelantó en la introducción (consúltense secciones 1.2.1 y 1.3.1) a pesar de que existe una gran diversidad de modos de selección, la mayoría de la investigación se ha centrado en el desarrollo de pruebas estadísticas para detectar la selección positiva. En este estudio hemos querido desarrollar dos intuitivos estimadores puntuales de la selección negativa o purificadora, el modo de selección más común, pero no por ello menos importante. La selección negativa consiste en la eliminación no al azar de un alelo debido a sus efectos perjudiciales sobre la eficacia biológica del individuo portador. La cantidad de evidencias en distintas especies a favor de la acción de la selección negativa y sus efectos sobre la variación neutra ligada es tal que muchos genéticos de poblaciones proponen incorporar la selección de fondo (Charlesworth 1993) a las teorías neutralistas (Kimura 1969a; 1983; Ohta y Kimura 1971) para obtener un modelo nulo actualizado que consiga mejorar nuestra interpretación de los patrones de variación genética a lo largo de los cromosomas y la detección de las regiones o genes del genoma que están detrás de la adaptación de las especies (Reed *et al.* 2005; McVicker *et al.* 2009; Hernandez *et al.* 2011; Lohmueller *et al.* 2011; Chun y Fay 2011; Charlesworth 2012a; Comeron 2014; Elyashiv *et al.* 2014; Corbett-Detig *et al.* 2015). De este modo, embarcarse en el diseño y aplicación de estimadores de la intensidad de la selección purificadora que sean por un lado de fácil implementación y por otro robustos a perturbaciones del equilibrio poblacional, como cambios demográficos recientes o la misma selección ligada, es más necesario que nunca.

En este estudio hemos propuesto dos estimadores o estadísticos que recogen distintos aspectos de la información contenida en el espectro de frecuencias alélicas, y por extensión en la *DFE* de las nuevas mutaciones deletéreas. El primero de estos estimadores,  $d_n$ , mide la proporción de sitios segregantes selectivos respecto a los sitios segregantes neutros por sitio selectivo y neutro, respectivamente. Es una

medida del constreñimiento en la fase polimórfica. Hemos mostrado una altísima correlación entre  $d_n$  y un sofisticado estimador cuantitativo de la proporción de mutaciones fuertemente deletéreas (con  $N_{es} < -10$ ) que tiene en cuenta las posibles alteraciones del espectro de frecuencias producidas por cambios demográficos recientes (Keightley y Eyre-Walker 2007). Mediante simulaciones hemos demostrado que  $d_n$  es muy robusto a cambios demográficos recientes, como el cuello de botella severo que se cree han sufrido las poblaciones no-africanas de *D. melanogaster* (Begun y Aquadro 1993; Andolfatto 2001; Li y Stephan 2006; Thornton y Andolfatto 2006). Además, el valor del estadístico  $d_n$  disminuye cuando la tasa de recombinación por sitio es menor a la tasa de mutación por sitio (véase figura 3.3). Este hecho indica que la interferencia entre alelos seleccionados (el denominado efecto Hill-Robertson, consultese sección 1.1.3) disminuye, como es de esperar, la eficacia de la selección negativa y con ello el constreñimiento en la fase polimórfica permitiendo que incluso alelos fuertemente deletéreos lleguen a segregar a una frecuencia apreciable en la población. Al igual que las estimas de la fracción de mutaciones fuertemente seleccionadas (con  $N_{es} < -10$ ) basadas en el método de Keightley y Eyre-Walker (2007), para todas nuestras estimas de  $d_n$  (es decir, en todas las ventanas cromosómicas y clases funcionales analizadas) la hipótesis nula es refutada y la hipótesis alternativa donde la selección negativa fuerte está actuando es aceptada (consultese tabla S7 del ANEXO). En definitiva, a pesar que el valor de  $d_n$  depende del tamaño de muestra, en la práctica las conclusiones extraídas con  $d_n$  y el estimador de la fracción de nuevas mutaciones fuertemente deletéreas de Keightley y Eyre-Walker (2007) son las mismas (véanse tabla 3.2 y 3.3). Con la ventaja que  $d_n$  es mucho más fácil y rápido de calcular. Aconsejamos el uso de  $d_n$  como una primera aproximación para cartografiar la intensidad de la selección purificadora a lo largo del genoma. No obstante, los resultados obtenidos con el método de Keightley y Eyre-Walker (2007) son más fáciles de comparar con estudios previos en *Drosophila* y estudios realizados en otras especies.

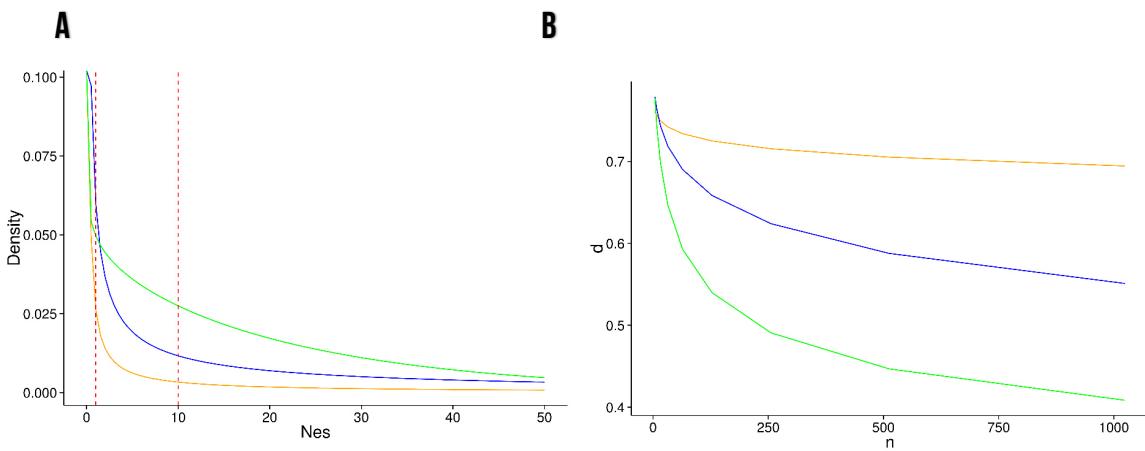
El segundo de nuestros estadísticos,  $b$ , es una medida del exceso de alelos selectivos segregando a baja frecuencia (por debajo del 5%) respecto a lo observado en sitios putativamente neutros. De nuevo la correlación entre  $b$  y un estimador cuantitativo de la proporción de mutaciones ligeramente deletéreas ( $-10 < N_e S < -1$ ) (Keightley y Eyre-Walker 2007) es muy alta. Sin embargo, nuestras simulaciones han demostrado que  $b$  es profundamente sensible a la demografía (véase figura 3.3); llegando a sobreestimar en más de un 40% el valor esperado bajo un tamaño de población constante. Esto es debido a que el espectro de frecuencias de mutaciones neutras esperado bajo el equilibrio mutación-deriva presenta una mayor proporción de alelos a frecuencias intermedias que el espectro de frecuencias de mutaciones seleccionadas esperado bajo el equilibrio mutación-selección-deriva. En otras palabras, después de un cuello de botella reciente las mutaciones seleccionadas alcanzan antes su frecuencia en el equilibrio (su distribución estacionaria) que las mutaciones neutras. Esta dinámica poblacional distinta para mutaciones neutras y seleccionadas es la responsable de dicha sobreestima de  $b$ , y aunque los valores significativos de  $b$  a lo largo del genoma son indicativos de que una fracción importante de los alelos está bajo selección negativa débil en *D. melanogaster* desaconsejamos el uso de dicho estadístico. Además, el estimador  $b$  no es capaz de capturar gran parte de los alelos ligeramente deletéreos que segregan en una muestra (lo que implica que muchos alelos ligeramente deletéreos segregan  $> 5\%$ ), ya que siempre genera estimas menores que el estadístico de Keightley y Eyre-Walker (2007) para nuevas mutaciones ligeramente deletéreas.

## ¿NUEVO ESTIMADOR CUANTITATIVO DE LA DFE DE NUEVAS MUTACIONES DELETÉREAS ROBUSTO A LA DEMOGRAFÍA?

Como se acaba de comentar, el estadístico  $d_n$  es muy robusto ante situaciones que puedan alterar el espectro de frecuencias. Nuestras simulaciones indican que la razón entre el número de sitios segregantes seleccionados y neutros (por sitio) alcanza su valor esperado bajo un tamaño de población constante y con libre

recombinación sorprendentemente rápido (nótese que el cuello de botella de *D. melanogaster* que hemos simulado se cree finalizó hace menos de  $0,0042N_e$  generaciones o 50 años [Thornton y Andolfatto 2006]). Esto significa que  $d_n$  alcanza su valor esperado bajo el equilibrio muy pocas generaciones después de intensas perturbaciones demográficas.

Nuestros resultados muestran un interesante potencial del estadístico  $d_n$ , el cual podría utilizarse para tratar de estimar cuantitativamente la DFE sin necesidad de calcular ni el SFS ni inferir la demografía de la especie. Pero ¿en qué consistiría dicho método? Véase figura 4.1 para un caso práctico. Se espera una relación negativa entre  $d_n$  y el tamaño de muestra, ya que, la probabilidad de que un alelo deletéreo segregue respecto a un alelo neutro depende del tamaño de muestra ( $n$ ) (consúltese sección 3.1.1 para más detalles). A su vez, la relación entre  $n$  y  $d_n$  dependerá fundamentalmente de la DFE para las nuevas mutaciones que ocurren en un *locus* de tamaño  $L$  con una tasa de mutación  $\mu$ . Podríamos tratar de implementar un método de computación bayesiana aproximada (ABC, *Approximate Bayesian Computation*, véase sección 1.3.3) donde distintas DFEs son simuladas y el estadístico  $d_n$  es estimado a distintos tamaños de muestra, después ajustaríamos una función potencial a la relación entre  $n$  y  $d_n$ :  $d_n = a n^{-b}$ , donde los parámetros  $a$  y  $b$  actuarían como nuestros estadísticos resumen (*summary statistics*) (análisis preliminares señalan que la función potencial es una muy buena aproximación, ya que el  $R^2 > 0.98$ , datos no mostrados). Esta es, en definitiva, una muy interesante línea de investigación que merece ser explorada en un futuro cercano. No obstante, es necesario simular más escenarios demográficos, DFEs y tamaños de muestra con tal de conocer hasta qué punto el valor de  $d_n$  es independiente, o no, de la historia demográfica de las especies. Sería muy necesario añadir también arrastres selectivos positivos recurrentes a las simulaciones, esto es el *genetic draft* (nótese que nuestras simulaciones sólo contemplaban el efecto de la selección de fondo [Charlesworth 1993]) y comprobar de nuevo el comportamiento de  $d_n$ .



**FIGURA 4.1** ¿Puede la relación entre  $d_n$  y el tamaño de muestra ( $n$ ) informarnos sobre la DFE subyacente? (A) Ejemplos de tres DFEs con distintos efectos promedio de las nuevas mutaciones sobre la eficacia biológica ( $\overline{N_e S}$ ) y valores del parámetro de forma ( $\beta$ ) de la distribución gamma asumida. Las líneas verticales discontinuas rojas delimitan por un lado la frontera entre las mutaciones efectivamente neutras y las efectivamente seleccionadas ( $|N_e S| = 1$ ) y por otro la frontera entre las mutaciones débilmente seleccionadas y fuertemente seleccionadas ( $|N_e S| = 10$ ). Nótese que, aunque se estén mostrando valores positivos de la DFE, en realidad se trata de la DFE de nuevas mutaciones deletéreas y efectivamente neutras, por lo tanto, la imagen debería girarse 180 grados. En **naranja** tenemos una DFE con  $\beta = 0,1$  y  $\overline{N_e S} = -1.000.000$ , esta DFE es muy representativa de la teoría neutralista original (Kimura 1969a), donde las mutaciones son o bien neutras o bien muy deletéreas. En **verde** ( $\beta = 0,9$  y  $\overline{N_e S} = -25$ ) tenemos una DFE con una gran proporción de mutaciones ligeramente deletéreas. En **azul** ( $\beta = 0,3$  y  $\overline{N_e S} = -300$ ) se muestra una DFE intermedia entre las dos anteriores. (B) Relación entre el valor esperado del estadístico  $d_n$  estimado con infinitos sitios segregantes y el tamaño de muestra ( $n$ ) para las tres DFEs mostradas en (A). Para  $n = 4$  las 3 DFEs tienen el mismo valor de  $d_n$  ( $d_n \sim 0,77$ ), es decir, aunque se trata de DFEs muy distintas son capaces de generar el mismo valor de  $d_n$ . Sin embargo, al aumentar  $n$  somos capaces de detectar diferencias importantes entre DFEs.

## 4.2 DFE DE LAS NUEVAS MUTACIONES DELETÉREAS Y TASA DE EVOLUCIÓN ADAPTIVA EN EL GENOMA DE *D. melanogaster*

Desde el comienzo de este trabajo de tesis ya existían buenos estimadores cuantitativos de la *DFE* (Keightley y Eyre-Walker 2007), los cuales utilizan la información contenida en el espectro de frecuencias de sitios putativamente neutros para estimar la demografía y dar así estimas corregidas de la *DFE* de nuevas mutaciones deletéreas a partir del espectro de frecuencias de sitios putativamente seleccionados. A partir de la *DFE* y utilizando la expresión [1.2] de Kimura (1957, 1983) se puede calcular el número esperado de substituciones neutras y ligeramente deletéreas que han ocurrido en los sitios putativamente seleccionados (Eyre-Walker y Keightley 2009). La diferencia entre el número de substituciones esperadas y el número de substituciones observadas corresponde a las sustituciones adaptativas (consúltese expresión [2.14]). Sin embargo, esta metodología no informa sobre el coeficiente de selección de las sustituciones adaptativas, sólo de la tasa de evolución adaptativa.

En este estudio hemos estimado la *DFE* y la tasa de evolución adaptativa a lo largo del genoma codificador y no-codificador de *D. melanogaster* distinguiendo entre brazos cromosómicos y contextos recombinacionales. Hacer esto nos ha permitido tratar de dar respuesta a estas cuatro cuestiones: (1) ¿Qué proporción del genoma es funcional? (2) ¿Cuál es la contribución relativa del genoma codificador y no-codificador a la variación genética de la eficacia biológica? (3) ¿Cuál es la contribución relativa del genoma codificador y no-codificador a la evolución adaptativa? (4) ¿Existen diferencias en la *DFE* y la tasa de evolución adaptativa entre brazos cromosómicos? ¿Si es así a qué se deben?

Sabemos que un 15% del genoma eucromático de *D. melanogaster* codifica para proteínas, un 3,2% es UTR, un 30% pertenece a secuencia intrónica y un 51,8% intergénica (Adams *et al.* 2000; Misra *et al.* 2002; Alexander *et al.* 2010,

[www.flybase.org](http://www.flybase.org)), sin embargo, hasta hace relativamente poco el grado de funcionalidad del genoma no-codificador de muchas especies se desconocía. Actualmente sabemos que la mayoría de genomas de eucariotas superiores está formado principalmente de ADN no-codificador, el cual debe contener la información necesaria para regular los niveles de expresión, el *tempo*, y la organización espacial de cientos o miles de genes codificadores, así como la orquestación de la replicación o el empaquetamiento de los cromosomas (Lewin 2007).

#### 4.2.1 FRACCIÓN FUNCIONAL DEL GENOMA DE *D. melanogaster*

Estimar qué proporción del genoma está sometido a la selección negativa es equivalente a responder la pregunta: ¿Qué proporción del genoma es funcional? Salvo algunas excepciones (véase más abajo). Estudios previos en *Drosophila*, y otras especies, que trataban de responder esta misma pregunta se basaban en aproximaciones basadas en datos filogenéticos (Bergman y Kreitman 2001; Andolfatto 2005; Siepel *et al.* 2005; Halligan y Keightley 2006), es decir, en estimas del grado de conservación de la secuencia o constreñimiento funcional entre parejas o grupos de especies más o menos cercanas evolutivamente (consultese sección 1.1.2). Este tipo de aproximaciones pueden subestimar la proporción de sitios funcionales si las substituciones beneficiosas son abundantes entre especies o si la clase de sitios neutra (tradicionalmente las posiciones sinónimas) están sometidas a selección negativa. Además, si existen dinámicas mutacionales distintas entre los sitios que asumimos evolucionan de forma neutra y los sitios seleccionados, esto puede ocurrir cuando la composición de bases es muy distinta entre clases de sitios, podrían producirse también estimas incorrectas del constreñimiento. Por último, el concepto "constreñimiento" se suele interpretar como una propiedad de los sitios en el genoma, en lugar de una propiedad de las posibles mutaciones que se producen en estos sitios. Por ejemplo, es posible que una pequeña región del genoma carezca totalmente de función, pero si nuevas mutaciones crean sitios espurios de unión a

factores de transcripción que pueden alterar la expresión de genes, estos sitios parecerán estar sometidos a selección negativa aunque no sean funcionales (Hahn *et al.* 2003; Clop *et al.* 2006). Además, la falta de evidencias de constreñimiento puede ser engañosa acerca de la función, como sugiere la identificación de *enhancers* en el genoma humano los cuales no están conservados entre especies (Blow *et al.* 2010). Así, mientras el constreñimiento o la conservación de la secuencia puede ser una primera aproximación razonable para conocer el grado de funcionalidad de un genoma, su interpretación a veces puede ser complicada.

En este estudio, en cambio, hemos utilizado la información contenida en el espectro de frecuencias para estimar la *DFE* y con ello el % de genoma que es funcional (como en Sawyer *et al.* 1987; Akashi y Schaeffer 1997; Keightley y Eyre-Walker 2007; Boyko *et al.* 2008). Hemos comparado el espectro de frecuencias de sitios putativamente neutros (en nuestro caso las posiciones 4-veces degeneradas) con el espectro de frecuencias de sitios putativamente seleccionados: sitios 0-veces degenerados, UTRs, intrones y región intergénica. Esta aproximación se fundamenta en el hecho de que la selección purificadora reduce la frecuencia a la cual segregan las mutaciones seleccionadas respecto a las mutaciones neutras. Esta aproximación tiene la ventaja que no se ve afectada por las substituciones adaptativas (las cuales son funcionales pero disminuyen la conservación de la secuencia entre especies) y que es muy robusta a diferencias en las tasas de mutación entre tipos de sitios, siempre y cuando, como hemos hecho en este estudio, no utilicemos el estado ancestral-derivado de las mutaciones (Tajima 1989; Hernandez *et al.* 2007; Baudry y Depaulis 2003). Además, es capaz de cuantificar aquellas regiones sometidas a selección negativa recientemente, dentro de un mismo linaje. Para una revisión de ambas metodologías y sus limitaciones consultese Zhen y Andolfatto (2013). Más allá de la aproximación utilizada nuestros resultados son los más completos hasta la fecha en cuanto a % de genoma analizado, calidad de la secuencia y número de individuos utilizados para estimar el espectro de frecuencias.

Observamos que globalmente, sin distinguir entre tipos de sitios, y sumando el número de mutaciones ligeramente deletéreas ( $-10 < N_{eS} < -1$ ) y fuertemente deletéreas ( $N_{eS} < -10$ ), el ~60% del genoma de *D. melanogaster* parece ser funcional, el resto (~40%) es efectivamente neutro ( $-1 < N_{eS} < 1$ ). Esta estima es cercana a la encontrada por Halligan y Keightley (2006) y Siepel *et al.* (2005), donde ~50% de los sitios entre *D. melanogaster* y especies cercanas parecen estar conservados. Nuestra estima es un 10% superior debido probablemente a las substituciones adaptativas las cuales son funcionales (están constreñidas en la fase polimórfica actualmente) pero disminuyen el constreñimiento entre especies. Estudios previos en *Drosophila* que utilizaban los niveles de divergencia en regiones intergénicas e intrones largos indicaban que >50% de este tipo de sitios están bajo selección purificadora (Bergman y Kreitman 2001; Andolfatto 2005; Bachtrog y Andolfatto 2006; Halligan y Keightley 2006; *Drosophila* 12 Genomes Consortium 2007), hecho que ha sido respaldado en análisis que incorporan el polimorfismo (Andolfatto 2005; Bachtrog y Andolfatto 2006; Casillas *et al.* 2007; Haddrill *et al.* 2008). En nuestro caso estimamos que el ~51% de las posiciones en regiones intergénicas e intrones parecen estar sometidas a selección negativa (véase tabla 3.2). No hemos observado diferencias en la proporción de mutaciones negativamente seleccionadas entre intrones y regiones intergénicas al igual que estudios previos (Bergman y Kreitman 2001; Andolfatto 2005; Halligan y Keightley 2006). El mayor constreñimiento en las secuencias UTR respecto al resto de ADN no-codificador ya había sido descrito con anterioridad (Andolfatto 2005); estimas previas del constreñimiento muestran que el ~60% de los sitios UTR son funcionales (véase tabla 2 de Andolfatto 2005). En nuestro conjunto de datos el 68% de los sitios UTRs están sometidos a selección (véase tabla 3.2), de nuevo una estima ligeramente superior a estimas previas. Finalmente, nótese que nuestro análisis aporta un nuevo dato: el genoma no-codificador es el que domina en lo que respecta a la proporción de nuevas mutaciones ligeramente deletéreas, donde un 21% de las nuevas mutaciones no-codificadoras (que ocurren tanto en UTR, intrón o región intergénica) están débilmente seleccionadas en comparación con el 9% estimado para las secuencias codificadoras.

Los patrones observados de constreñimiento en el ADN no-codificador de *D. melanogaster* son muy distintos al encontrado en ratones, donde el constreñimiento se desvanece o alcanza un valor muy bajo a < 3 kb de distancia de la secuencia codificadora tanto en región intergénica como en intrones (Gaffney y Keightley 2006). En homínidos las diferencias son incluso mayores, donde el constreñimiento promedio alrededor de las secuencias codificadoras es cercano a cero (Keightley *et al.* 2005a; 2005b; Kryukov *et al.* 2005). Estas diferencias entre mamíferos y *Drosophila* en cuanto a la proporción de ADN no-codificador funcional podría deberse al mayor  $N_e$  de *Drosophila* y/o al menor tamaño del genoma de *Drosophila* respecto estos taxones (Halligan y Keightley 2006), el cual es un orden de magnitud menor – el genoma de *Drosophila* es muy compacto. El hecho que hayamos encontrando un elevado constreñimiento en las regiones intergénicas de hasta 5 kb aguas arriba y aguas abajo del primer y último exón, respectivamente, sugiere que los sitios funcionales (potencialmente reguladores de la transcripción) no sólo se concentran cerca de los genes como cabría esperar *a priori*. En el estudio de Andolfatto (2005) se encuentra, de hecho, que cerca de los genes el constreñimiento ( $C$ ) es ligeramente inferior: en regiones intergénicas a menos de 2 kb de distancia del gen más cercano  $C = 40,6\%$  y a más de 4 kb de distancia del gen más cercano  $C = 54,6\%$ . El trabajo de Nelson *et al.* (2004) da una explicación funcional a este mayor constreñimiento lejos de los genes. Nelson *et al.* (2004) muestra que los genes con patrones de expresión complejos tienden a estar asociados a secuencias intergénicas largas las cuales son ricas en elementos reguladores. En cualquier caso, el elevado constreñimiento funcional del genoma no-codificador de *D. melanogaster* podría estar mediado por un elevado número de elementos reguladores (Bergman *et al.* 2002; Emberly *et al.* 2003; Sironi *et al.* 2005) y/o a la presencia de genes codificadores (Storz 2002) y no-codificadores (Stolc *et al.* 2004) no anotados. Aunque con la incorporación de las anotaciones de modENCODE (The modENCODE Project Consortium 2010) al último archivo de anotaciones esta última posibilidad es cada vez menos probable.

Nuestras estimas de la *DFE* para nuevas mutaciones deletéreas coinciden en general con estimas realizadas en estudios previos, si bien, hasta la fecha en *Drosophila* solamente se disponía de estimas de la *DFE* para mutaciones de cambio de aminoácido (Eyre-Walker *et al.* 2002; Loewe *et al.* 2006; Loewe y Charlesworth 2006; Eyre-Walker y Keightley 2009; Andolfatto *et al.* 2011; Wilson *et al.* 2011), en intrones (Eyre-Walker y Keightley 2009) y en secuencias no-codificadoras muy conservadas (Casillas *et al.* 2007), para el resto de clases funcionales disponemos únicamente de estimas de constreñimiento las cuales no informan sobre los detalles de la *DFE* como la proporción de mutaciones ligeramente deletéreas. El trabajo de Eyre-Walker y Keightley (2009) es el más similar metodológicamente al nuestro y está basado en 12 individuos de una población africana de *D. melanogaster* y 688 loci en total (397 genes codificadores y 291 intrones). Nuestras estimas indican que ~51% de las posiciones intrónicas son funcionales (como Bergman y Kreitman 2001; Andolfatto 2005; Bachtrog y Andolfatto 2006; Halligan y Keightley 2006; *Drosophila* 12 Genomes Consortium 2007), mientras que Eyre-Walker y Keightley (2009) estimaron una cantidad menor, de alrededor del 30%. Estas diferencias podrían deberse a un sesgo en la muestra de Eyre-Walker y Keightley (2009) hacia intrones pequeños (< 80-90 pb), los cuales se han descrito evolucionando neutralmente en *Drosophila* (Halligan y Keightley 2006; Parsch *et al.* 2010). De hecho, este parece ser el caso, los intrones de Eyre-Walker y Keightley (2009) son en promedio de ~200 pb (véase tabla 2 de Eyre-Walker y Keightley 2009), mientras el valor genómico promedio es de ~1,5 kb (Zhu *et al.* 2009). Nuestras estimas de la *DFE* son más cercanas al valor genómico pues no hemos filtrado ni categorizado los intrones por su tamaño. Las estimas previas de la *DFE* para nuevas mutaciones no-sinónimas en *Drosophila* (Eyre-Walker *et al.* 2002; Loewe *et al.* 2006; Loewe y Charlesworth 2006; Eyre-Walker y Keightley 2009; Andolfatto *et al.* 2011; Wilson *et al.* 2011) son cualitativamente equivalentes a las encontradas por nosotros en este trabajo, donde menos de un 10% de las nuevas mutaciones no-sinónimas son efectivamente neutras, un 80% son fuertemente deletéreas y alrededor de un 10% son ligeramente deletéreas. En conclusión, en

general nuestros resultados respaldan los resultados encontrados en trabajos anteriores, pero el hecho que nuestras estimas sean a nivel genómico, se hayan calculado a partir de una gran muestra de individuos y no estén basadas en muestras fragmentarias del genoma las convierte en las estimas de referencia en *D. melanogaster*.

#### 4.2.2 CONTRIBUCIÓN DE LAS MUTACIONES CODIFICADORAS Y NO-CODIFICADORAS A LA VARIACIÓN EN LA EFICACIA BIOLÓGICA ENTRE INDIVIDUOS DE *D. melanogaster*

Conocer la *DFE* de nuevas mutaciones deletéreas para secuencias codificadoras y no-codificadoras no tan solo nos permite saber cuál es la fracción funcional del genoma de *D. melanogaster* (es decir, con  $N_e s < -1$ ), sino saber cuál es también la contribución relativa del genoma codificador y no-codificador a la variación en la eficacia biológica. Bajo un modelo de equilibrio mutación-selección, la varianza genética de la eficacia biológica en el equilibrio es proporcional al producto del efecto promedio de las nuevas mutaciones deletéreas sobre la eficacia biológica ( $\overline{N_e s}$ ) y la tasa de mutación genómica ( $\mu$ ) (Eyre-Walker 2010). Por lo tanto, las mutaciones que contribuyen más a la varianza en la *fitness* son aquellas muy numerosas y/o que están más fuertemente seleccionadas (Eyre-Walker 2010). No obstante, las estimas de  $\overline{N_e s}$  para las nuevas mutaciones deletéreas dependen de la frecuencia de mutaciones muy fuertemente seleccionadas, las cuales es poco probable que segreguen en una muestra como la nuestra de  $n = 128$ . Serían necesarias muestras de 1.000, 10.000 o millones de individuos para obtener estimas más precisas del  $\overline{N_e s}$  (Eyre-Walker y Keightley 2009; Eyre-Walker 2010; Keightley y Eyre-Walker 2010; Halligan *et al.* 2013). Si pretendemos comparar la intensidad o fuerza de la selección entre categorías de sitios, trabajos previos han mostrado que la proporción de mutaciones en distintos intervalos de la *DFE* es una estima más robusta que el efecto promedio de las nuevas mutaciones (Keightley y Eyre-Walker 2007; Kousathanas y Keightley 2013). Estos intervalos son los que mostramos en las tablas 3.2-3.4, sin

embargo, y aunque deben tratarse con mucha precaución hemos encontrado que nuestra estima promedio del  $\overline{N_eS}$  para mutaciones no-sinónimas es mucho mayor que nuestras estimas promedio del  $\overline{N_eS}$  para cualquier clase de sitio no-codificador (sitios 0-veces degenerados  $\overline{N_eS} \sim -15.000$ , UTR  $\overline{N_eS} \sim -1.000$ , intrones  $\overline{N_eS} \sim -20$  y regiones intergénicas  $\overline{N_eS} \sim -17$ ). En resumen, asumiendo la misma tasa de mutación en el ADN codificador y no-codificador, y ponderando por la longitud relativa del ADN codificador y los distintos tipos de ADN no-codificador, podemos concluir que en el ADN codificador se encuentra la mayoría de la variación genética que está detrás de las diferencias en la eficacia biológica entre individuos de *Drosophila*, como también parece ser el caso en ratones (Halligan *et al.* 2013) y humanos (Eyre-Walker 2010). Sin embargo, no podemos dar una estima precisa del % de la variación en la *fitness* que explica el ADN codificador respecto al resto de categorías no-codificadoras porque, como se ha dicho, las estimas de  $N_eS$  son poco precisas bajo nuestro tamaño de muestra (para una revisión de este problema consultese Keightley y Eyre-Walker 2010).

Nuestros resultados indican que a pesar que las secuencias codificadoras son minoría en el genoma, el conjunto de mutaciones que allí ocurren son más deletéreas que el conjunto de mutaciones que ocurren fuera. Nótese, no obstante, que hasta ahora nuestra definición de ADN no-codificador no distingue entre ADN no-codificador conservado y no-conservado, cuando sabemos que un ~20–30% del ADN no-codificador está muy conservado entre especies de insectos (Bergman y Kreitman 2001; Bergman *et al.* 2002; Siepel *et al.* 2005). Casillas *et al.* (2007) mostró que esta conservación se debe a la acción de la selección purificadora y estimó la *DFE* para mutaciones deletéreas en secuencias CNS (*conserved noncoding sequences*) del cromosoma X. Estimaron un  $\overline{N_eS} = -30,7$  asumiendo, como nosotros, una distribución gamma para modelar la *DFE*. Este valor es casi 500 veces menor al observado para los cambios no-sinónimos (véase más arriba) aunque sólo hay ~5,7 veces más ADN no-codificador (conservado y no-conservado) que codificador. Por lo

tanto, de nuevo, parece muy probable que los cambios estructurales en proteínas dominen las diferencias en la *fitness* incluso en relación al ADN no-codificador muy conservado.

Finalmente, nótese que sólo hemos discutido aquí el efecto de mutaciones puntuales, sin embargo, las inserciones, delecciones, inversiones y otras reorganizaciones cromosómicas están involucradas muchas veces en graves enfermedades humanas (Lupski 2009). Este tipo de mutaciones, aunque mucho menos numerosas que las mutaciones puntuales, podrían contribuir mucho a la varianza en la *fitness* porque sus efectos sobre esta serán probablemente mucho mayores.

#### 4.2.3 CONTRIBUCIÓN DE LAS SUBSTITUCIONES CODIFICADORAS Y NO-CODIFICADORAS A LA EVOLUCIÓN ADAPTATIVA EN *Drosophila*

La pregunta sobre cuál es la contribución relativa del genoma codificador y el genoma no-codificador a la evolución adaptativa no es nueva. King y Wilson (1975) propusieron que los cambios en las proteínas (denominados cambios estructurales) difícilmente podrían explicar la miríada de adaptaciones fenotípicas que diferencia a humanos de chimpancés, o viceversa. Para ellos los cambios reguladores en secuencias no-codificadoras que afectan al *timing* y la especificidad de la expresión génica deberían dominar el cambio evolutivo adaptativo. Muchos estudios han propuesto que la evolución de la regulación de la expresión génica podría estar detrás de la organización modular, la diversificación funcional y el origen de nuevos caracteres en organismos superiores (Stern 2010; Wray *et al.* 2003; Davidson 2001; Carroll 2000). Existen evidencias empíricas de experimentos de cartografía genética (*genetic mapping*) que apoyan (Carrol 2005; Wray 2007; Jones *et al.* 2012) y contradicen esta hipótesis (Hoekstra y Coyne 2007). Desde un punto de vista teórico se ha discutido que los cambios reguladores deberían causar menores efectos pleiotrópicos perjudiciales que las mutaciones que afectan a proteínas (Carrol 2005).

Se han encontrado evidencias de selección positiva tanto en el ADN no-codificador de *Drosophila* y ratones (Andolfatto 2005; Torgerson *et al.* 2009; Eyre-Walker y Keightley 2009; Kousathanas *et al.* 2011; Halligan *et al.* 2011; Halligan *et al.* 2013) como en el ADN codificador de *Drosophila*, ratones, bacterias y algunas plantas (Bustamante *et al.* 2002; Smith y Eyre-Walker 2002; Bierne y Eyre-Walker 2003; Sawyer *et al.* 2003; Charlesworth y Eyre-Walker 2006; Haddrill *et al.* 2010; Ingvarsson 2010; Slotte *et al.* 2010; Strasburg *et al.* 2011), a su vez hay escasas evidencias de selección positiva en genes codificadores de homínidos y algunas plantas (Chimpanzee Sequencing and Analysis Consortium 2005; Zhang y Li 2005; Boyko *et al.* 2008; Eyre-Walker y Keightley 2009; Gossmann *et al.* 2010). Nos preguntamos qué tipo de cambios, estructurales o reguladores, son más importantes para la adaptación en *Drosophila*.

La tabla 4.1 recoge estimas previas de la fracción de substituciones adaptativas ( $\alpha$ ) en comparación con nuestras estimas en el genoma codificador y no-codificador de *D. melanogaster*. En general, nuestras estimas corroboran los resultados a los que apuntaban estudios previos, las evidencias de selección positiva son claras tanto en el genoma codificador como no-codificador de *Drosophila*. Estos valores de  $\alpha$  positivos en todas las clases de sitios podrían interpretarse alternativamente como un menor  $N_e$  en el pasado (que llevó a la fijación de mutaciones ligeramente deletéreas) respecto al presente (donde estas mutaciones ya no segregan debido al aumento de la eficacia de la selección) (Nei *et al.* 2010; Fay 2011) (sección 1.3.2). Sin embargo, las elevadas estimas de  $\alpha$  entre distintas especies de *Drosophila* (véase el grupo *D. americana* [Maside y Charlesworth 2007] y la especie *D. miranda* [Bachtrog 2008]) nos hace decantarnos por una elevada tasa de adaptación general en *Drosophila* y no tanto por un aumento reciente del censo efectivo en todas las especies del género *Drosophila* estudiadas.

**TABLA 4.1 ESTIMAS DE  $\alpha$  EN *D. melanogaster* EN ESTUDIOS PREVIOS Y ESTE ESTUDIO**

		Previous estimates	This study
Non-synonymous	A	30%, n = 419, <i>D. simulans</i> (Shapiro et al. 2007)	
		52%, n = 397, <i>D. simulans</i> (Eyre-Walker y Keightley 2009)	
		12%, n = ?, <i>D. simulans</i> (Langley et al. 2012)	40%
	X	24%, genomic, <i>D. yakuba</i> (Mackay et al. 2012)	
		86%, n = 105, <i>D. simulans</i> (Andolfatto et al. 2011)	
		46%, n = ?, <i>D. simulans</i> (Langley et al. 2012)	53%
UTR	A	47%, genomic, <i>D. yakuba</i> (Mackay et al. 2012)	
		6%, genomic, <i>D. yakuba</i> (Mackay et al. 2012)	32%
	X	20%, genomic, <i>D. yakuba</i> (Mackay et al. 2012)	
		56%, n = 31, <i>D. simulans</i> (Andolfatto 2005)	45%
Intron	A	29%, genomic, <i>D. yakuba</i> (Mackay et al. 2012)	
		23%, n = 291, <i>D. simulans</i> (Eyre-Walker y Keightley 2009)	20%
	X	35%, genomic, <i>D. yakuba</i> (Mackay et al. 2012)	
Intergenic	A	19%, n = 72, <i>D. simulans</i> (Andolfatto 2005)	
		28%, genomic, <i>D. yakuba</i> (Mackay et al. 2012)	34%
	X	37%, genomic, <i>D. yakuba</i> (Mackay et al. 2012)	
		49%, n = 50, <i>D. simulans</i> (Andolfatto 2005)	17%

La primera columna muestra la clase funcional analizada. A y X corresponde a los resultados en autosomas y el cromosoma X, respectivamente. n es el número de loci utilizados para realizar las estimas, el estudio de Mackay et al. 2012 fue el primero en dar estimas promedio a escala genómica. Se muestra el taxón externo (*D. simulans* o *D. yakuba*) utilizado para estimar la divergencia junto con la referencia del artículo.

Dicho esto, si asumimos que la tasa de evolución adaptativa está limitada por la entrada de nuevas mutaciones (tanto en regiones codificadoras como no-codificadoras), entonces la tasa de evolución adaptativa es proporcional al producto de la fracción de nuevas mutaciones que son adaptativas ( $p_a$ ) y su tasa de fijación una vez aparecen en la población (Ohta y Kimura 1971). Si el tamaño de población es grande (como lo es el de *D. melanogaster*), la tasa de fijación de nuevas mutaciones beneficiosas en una población panmíctica es proporcional al producto del censo efectivo,  $N_e$ , y su efecto promedio en el heterocigoto,  $s_a/2$  (Ohta y Kimura 1971). Para una tasa de fijación de mutaciones neutras y ligeramente deletéreas dada,  $\alpha$  estará controlado por los parámetros  $p_a$  y  $N_e s_a$ . La tasa de substituciones adaptativas por sitio putativamente seleccionado es el producto de estos dos parámetros ( $K_+ = p_a N_e s_a$ ).

En estudios relativamente recientes  $p_a$  y  $N_{eS_a}$  han sido estimados por separado a través de distintos métodos que utilizan datos de polimorfismo y divergencia en *Drosophila*. Se han obtenido resultados contradictorios, estimas de una baja  $p_a$  pero considerable  $N_{eS_a}$ , es decir, de pocas mutaciones adaptativas pero fuertemente seleccionadas (Eyre-Walker 2006; Li y Stephan 2006; Macpherson *et al.* 2007; Jensen *et al.* 2008; Jensen 2009), así como de escenarios con muchas mutaciones adaptativas (elevada  $p_a$ ) pero débilmente seleccionadas (bajo  $N_{eS_a}$ ) (Sawyer *et al.* 2003; Andolfatto 2007; Schneider *et al.* 2011) e incluso una mezcla de mutaciones fuertemente y débilmente seleccionadas (Sattath *et al.* 2011; Wilson *et al.* 2011). Independientemente del valor verdadero de los parámetros  $p_a$  y  $N_{eS_a}$ , se conocen tres fuentes de información/metodologías que podrían informarnos sobre la DFE o como mínimo del  $\overline{N_eS}$  de las nuevas mutaciones adaptativas: (1) el decaimiento de la diversidad genética neutra alrededor de substituciones en sitios putativamente seleccionados (Sella *et al.* 2009; Lee *et al.* 2014), (2) el espectro de frecuencias de alelos derivados (Schneider *et al.* 2011) y (3) el exceso de variantes a alta frecuencia tras un arrastre selectivo positivo (Fay y Wu 2000). Las estimas más recientes de  $p_a$  y  $N_{eS_a}$  para mutaciones no-sinónimas en *D. melanogaster* indican que una proporción considerable ( $p_a \sim 1,5\%$ ) de las nuevas mutaciones no-sinónimas están débilmente seleccionadas a favor ( $N_{eS_a} \sim 5$ ) (Schneider *et al.* 2011). No hay estimas del valor de estos parámetros en el ADN no-codificador hasta la fecha.

En este estudio no hemos utilizado ninguna de estas aproximaciones para estimar los parámetros  $p_a$  y  $N_{eS_a}$  en el ADN codificador y no-codificador. Sin embargo, nuestros resultados corroboran indirectamente las estimas de Schneider *et al.* (2011) para el ADN codificador y pueden dar una idea aproximada del valor de estos parámetros para el ADN no-codificador. Mostramos que la relación entre la tasa de recombinación y nuestro estimador de la tasa de adaptación  $\omega_A$  ( $\omega_A = K_s/K_d$ , donde  $K_d$  es la tasa de substitución en las posiciones sinónimas 4-veces degeneradas y es nuestro indicador de la tasa de mutación neutra) es muy similar para secuencias codificadoras y no-

codificadoras (véase figura 3.7). Esta es una relación curvilínea – para valores bajos de recombinación ( $< 2 \text{ cM/Mb}$ ) existe una fuerte correlación positiva entre la tasa de adaptación y la tasa de recombinación, pero para valores de recombinación  $> 2 \text{ cM/Mb}$  no existe relación entre la adaptación y la recombinación. Esta relación sugiere que el parámetro  $N_e s_a$  podría ser muy similar para el ADN no-codificador y el ADN codificador en *Drosophila*. Si el parámetro  $N_e s_a$  fuese mayor a 10, o mayor a 100, es decir, si las nuevas mutaciones adaptativas estuviesen fuertemente seleccionadas, su probabilidad de ser adaptativas debería ser menor,  $p_a < 0.15\%$  o  $p_a < 0.015\%$ , respectivamente, con tal de mantener constante  $K_+$ . En este hipotético escenario donde las nuevas mutaciones beneficiosas fuesen poco comunes, pero estuviesen fuertemente seleccionadas, la interferencia de Hill-Robertson (Hill y Robertson 1966) afectaría muy poco a la probabilidad de fijación de las nuevas mutaciones adaptativas y la correlación entre adaptación y recombinación sería muy menor o inexistente. De hecho, este parece ser el caso para las substituciones adaptativas no-sinónimas en el cromosoma X, en este cromosoma no hay correlación entre la tasa de adaptación y la tasa de recombinación para los sitios no-sinónimos (Campos *et al.* 2014) (véase figura 3.7B). Hemos demostrado que esta falta de correlación no se debe a la menor densidad génica del cromosoma X respecto a los autosomas (véase figura 3.9), ni a una falta de poder estadístico debido al menor número de ventanas disponibles en el cromosoma X. Tampoco se debe a la perdida de resolución del mapa de recombinación, ya que, originalmente, realizamos el análisis en ventanas de 1 Mb mientras que las estimas de recombinación fueron estimadas sobre ventanas de 100 kb. Al repetir los análisis agrupando los genes por su tasa de recombinación y no por la ventana cromosómica a la que pertenecen seguimos observando una falta de correlación entre adaptación no-sinónima y recombinación en el X. Estas diferencias deben deberse a un  $N_e s_a$  mayor de las mutaciones no-sinónimas en el X que en los autosomas, lo cual las haría “insensibles” a la interferencia de Hill-Robertson.

En este trabajo hemos recopilado evidencias que indican que  $K_+$  es en promedio dos veces mayor en el ADN no-codificador que en el ADN codificador (véase tabla 3.2) – no hemos observado diferencias significativas en  $K_+$  entre distintos tipos de sitios no-codificadores. Teniendo en cuenta que hay 5,6 veces más ADN no-codificador que codificador, sólo el ~ 8% de las sustituciones adaptativas se han dado en el ADN codificador. No obstante, para la adaptación no sólo es importante considerar el número absoluto de sustituciones adaptativas sino el efecto promedio sobre la *fitness* de dichas sustituciones (es decir, el parámetro  $N_e s_a$ ), ya que, una categoría de sitios con una muy baja tasa de sustituciones adaptativas puede contribuir enormemente a la adaptación si estas sustituciones son fuertemente adaptativas (consultese expresión [2] de Halligan *et al.* 2013). En otras palabras, la mayor tasa de evolución adaptativa en el ADN no-codificador no implicaría, necesariamente, que las sustituciones adaptativas de las regiones no-codificadoras fuesen las dominadoras del cambio adaptativo en *Drosophila*. Sin embargo, si como sugiere la relación entre la tasa de adaptación y la tasa de recombinación la mayoría de mutaciones beneficiosas (codificadoras y no-codificadoras) están débilmente seleccionadas (por ejemplo, con  $N_e s_a \sim 5$  como sugiere Schneider *et al.* 2011), entonces el ADN no-codificador será seguramente el claro dominador de la evolución adaptativa en *Drosophila* (como ya sugirió Andolfatto 2005). Este también parece ser el caso en humanos según Enard *et al.* (2014). Dado que hay ~ 5,6 veces más sitios no-codificadores que codificadores y que la probabilidad que una nueva mutación sea beneficiosa ( $p_a$ ) será probablemente el doble en el ADN no-codificador que en el ADN codificador, entonces por cada sustitución adaptativa no-sinónima ocurren ~ 11,3 sustituciones adaptativas no-codificadoras en *D. melanogaster*.

Finalmente, como en el caso de la variación en la *fitness*, en este trabajo sólo hemos estudiado las mutaciones puntuales, si otro tipo de mutaciones estructurales (véase tabla 1.1) a pesar de tener tasas de mutación mucho menores, tuviesen un  $N_e s_a$  de tal magnitud que superara al de todas las sustituciones de un solo nucleótido juntas,

entonces no podríamos asegurar el dominio de las mutaciones (puntuales) no-codificadoras (y probablemente reguladoras de la expresión génica) sobre el cambio adaptativo. De hecho, nuestras conclusiones son sólo preliminares hasta que no tengamos mejores estimas de la *DFE* para otros tipos de mutaciones.

#### 4.2.4 EVOLUCIÓN LENTA DEL X: MUTACIONES DELETÉREAS RECESIVAS

La eficacia de la selección natural depende de una serie de parámetros como la *DFE*, la distribución de los coeficientes de dominancia fenotípica de las variantes alélicas y el censo efectivo. Cuando los efectos de la selección son relativamente débiles respecto a la deriva genética, el efecto Hill-Robertson (Hill y Robertson 1966) (véase sección 1.1.3) puede contribuir también a la variación en la eficacia de la selección. Este hecho hace que el contexto recombinacional y cromosómico, como la densidad génica y la tasa de mutación local, jueguen un importante papel en la tasa de fijación de alelos beneficiosos y alelos deletéreos a lo largo del genoma.

Multitud de estudios sugieren que la eficacia de la selección varía dentro de un mismo genoma. Por ejemplo, la razón entre el polimorfismo no-sinónimo y el polimorfismo sinónimo es mayor en autosomas que en el cromosoma X en *D. melanogaster* (Begun 1996; Andolfatto 2001; Mackay 2012; Campos *et al.* 2014). Este patrón puede deberse a una menor eficacia de la selección purificadora en los autosomas debido a la presencia de mutaciones deletéreas no-sinónimas (parcial o completamente) recesivas (Charlesworth *et al.* 1987), las cuales sí son visibles para la selección en machos (estos son hemicigóticos para el cromosoma X). Si las mutaciones beneficiosas son anecdóticas esperaríamos encontrar una menor tasa de evolución en el cromosoma X que en los autosomas; esta es la hipótesis del X lento (Charlesworth *et al.* 1987).

No obstante, el censo efectivo del cromosoma X es  $\frac{3}{4}$  el de los autosomas (si asumimos un censo efectivo de ambos sexos similar) lo que implicaría una menor eficacia de la selección en el X (Wright 1931). Por lo tanto, bajo el modelo más sencillo, la mayor o menor eficacia de la selección purificadora en el cromosoma X respecto los autosomas dependerá del balance de fuerzas entre la proporción de nuevas mutaciones deletéreas recesivas y el número de machos y hembras en la población. En la población putativamente ancestral de *D. melanogaster* del este de África el censo efectivo (estimado a partir de la diversidad nucleotídica,  $n$  [Tajima 1983], en sitios putativamente neutros) del cromosoma X respecto a los autosomas es igual a 1,1 (Andolfatto 2001; Hutter *et al.* 2007; Singh *et al.* 2007; Langley *et al.* 2012), igual que para la población ancestral africana de *D. simulans* (Andolfatto 2001; Singh *et al.* 2007). Esto no es debido a una mayor tasa de mutación en el cromosoma X – la tasa de mutación en X y autosomas es similar (véase Bauer y Aquadro 1997; Hutter *et al.* 2007; Keightley *et al.* 2009; Zeng y Charlesworth 2010; Haddrill *et al.* 2011; Hu *et al.* 2013). Trabajos recientes sugieren que esto podría ser debido a una menor tasa de entrecruzamiento en los autosomas respecto al cromosoma X, este hecho implicaría una mayor intensidad del efecto Hill-Robertson en estos (Charlesworth 2012b). La recombinación es más elevada en el cromosoma X porque los machos de *Drosophila* no pueden recombinar entre cromosomas homólogos (Ashburner *et al.* 2005); la tasa de recombinación efectiva para una tasa de recombinación  $r$  dada entre dos *loci* en hembras es  $\frac{1}{2}$  para los autosomas y  $\frac{2}{3}$  para el X (Charlesworth y Charlesworth 2010, p.381). En cambio, en las poblaciones no-africanas de *D. melanogaster* (y *D. simulans*), como la utilizada en este estudio, el censo efectivo del cromosoma X respecto a los autosomas es menor a  $\frac{3}{4}$  (alrededor de  $\frac{2}{3}$ ) (Aquadro *et al.* 1994; Begun y Aquadro 1993; Begun y Whitley 2000; Andolfatto 2001; Harr *et al.* 2002; Hutter *et al.* 2007; Singh *et al.* 2007; Stephan 2010; Mackay *et al.* 2012; Langley *et al.* 2012). Visto desde otro punto de vista, la razón entre la diversidad nucleotídica sinónima en poblaciones africanas y no-africanas es de  $\sim 1,35$  para autosomas y  $\sim 2,34$  para el cromosoma X (véase tabla 7 en Langley *et al.* 2012), es

decir, fuera de África la variación se ha reducido más en el X que en los autosomas. Dos posibles explicaciones, no mutuamente excluyentes, pueden explicar la mayor reducción del censo efectivo del cromosoma X respecto autosomas en poblaciones no-africanas: (1) mayor número de arrastres selectivos positivos promovidos por mutaciones beneficiosas recesivas en poblaciones que se enfrentan a nuevas condiciones ambientales (Charlesworth *et al.* 1987; Vicoso y Charlesworth 2009) y/o (2) cuellos de botella demográficos que afectarían más al cromosoma X que a los autosomas (Charlesworth 2001; Pool y Nielsen 2007; 2008).

En este estudio hemos encontrado diferencias significativas en la fracción global (sin distinguir entre ADN codificador y no-codificador) de nuevas mutaciones efectivamente neutras y fuertemente deletéreas entre el brazo cromosómico 3R y el cromosoma X (véase tabla 3.3). El brazo cromosómico 3R tiene más mutaciones efectivamente neutras (un 9%) y menos mutaciones fuertemente deletéreas (un 7%) que el cromosoma X (véase tabla 3.3) a pesar que el cromosoma X es el que tiene la menor densidad génica dentro de los 5 grandes brazos cromosómicos. No obstante, cuando comparamos regiones autosómicas y del X con tasas de recombinación y densidades génicas similares estas diferencias entre el brazo 3R y el cromosoma X se desvanecen lo cual sugiere que una proporción importante del brazo 3R tiene problemas para deshacerse de las mutaciones deletéreas (véase más abajo). Con este conjunto de datos emerge un nuevo patrón; el cromosoma X tiene una fracción significativamente menor de mutaciones ligeramente deletéreas no-sinónimas que los autosomas. Este resultado se ve respaldado por el conjunto de trabajos que sugieren que las nuevas mutaciones ligeramente deletéreas son parcialmente recesivas en *D. melanogaster* (con coeficiente de dominancia promedio  $h$  entre 0,1 y 0,4 dependiendo del estudio) (Mukai 1964; Crow y Simmons 1983; García-Dorado y Caballero 2000; revisado por García-Dorado *et al.* 2004). Si como parece este es el caso, entonces el hecho que los machos sean hemicigóticos para el X explicaría nuestros resultados. Es importante mencionar que no hemos detectado diferencias

significativas en la fracción de mutaciones no-codificadoras ligeramente deletéreas entre X y autosomas cuando comparamos regiones con densidades génicas y tasas de recombinación efectivas equivalentes (véase tabla 3.4). Esto implicaría que las mutaciones deletéreas que ocurren en regiones no-codificadoras tienen en promedio efectos aditivos ( $h = 0,5$ ). Sin embargo, trabajos teóricos muestran que a efectos prácticos la dinámica poblacional de mutaciones muy débilmente seleccionadas parcialmente recesivas se asemeja mucho al de mutaciones con efectos aditivos (Wright 1934; Kacser y Burns 1981; Charlesworth 2012b). Si una proporción importante de las nuevas mutaciones ligeramente deletéreas no-codificadoras se encuentran cerca de la frontera entre lo efectivamente neutro y lo efectivamente seleccionado ( $N_e s \sim -1$ ), el coeficiente de dominancia prácticamente no afectará a la dinámica poblacional de dichas mutaciones (por esta razón, quizás, no observamos diferencias entre X y autosomas). En cambio, las mutaciones no-sinónimas ligeramente deletéreas es más probable que se encuentran cerca de la frontera entre lo fuertemente deletéreo y lo ligeramente deletéreo ( $N_e s \sim -10$ ) y el coeficiente de dominancia sí afectará en este caso a la eficacia de la selección entre X y autosomas. Además, el menor  $N_e$  del cromosoma X respecto a los autosomas en poblaciones derivadas no-africanas como la nuestra (véase más arriba), puede dificultar más si cabe encontrar diferencias significativas para mutaciones muy débilmente seleccionadas.

Según el trabajo de Singh *et al.* (2008), el primero a escala genómica con un conjunto de datos de 6698 genes codificadores ortólogos uno a uno en 12 especies de *Drosophila*, la eficacia de la selección purificadora (medida por el mayor o menor uso de codones) es mayor en los genes del cromosoma X que en los genes autosómicos de las 12 especies de *Drosophila* estudiadas. Sin embargo, este trabajo no tenía en cuenta la mayor tasa de entrecruzamiento en el X respecto a los autosomas, ni su menor densidad génica, ambas propiedades disminuyen el efecto Hill-Robertson en el cromosoma X y podrían explicar estas diferencias (Charlesworth 2012b). En esta

tesis hemos comparado regiones autosómicas y del X que *a priori* están sometidas por igual al efecto Hill-Robertson; la menor fracción de mutaciones no-sinónimas ligeramente deletéreas en el X debe ser producto de la presencia de mutaciones deletéreas parcialmente recesivas en esta clase de sitios y no a una mayor eficacia de la selección negativa debido a la menor intensidad del efecto Hill-Robertson en el X.

Otro ejemplo que muestra que la eficacia de la selección purificadora varía a lo largo de los cromosomas es el siguiente. Regiones de muy baja recombinación en el genoma de *Drosophila* también muestran niveles más elevados de polimorfismo no-sinónimo, hecho que indicaría un menor  $N_e$  y eficacia de la selección negativa en dichas regiones (Presgraves 2005; Betancourt *et al.* 2009; Mackay *et al.* 2012; Campos *et al.* 2014). Respaldando este resultado, hemos encontrado una correlación negativa entre la fracción de nuevas mutaciones efectivamente neutras ( $-1 < N_e s < 1$ ) y la tasa de recombinación (tanto para ADN codificador como no-codificador en autosomas y cromosoma X) (véase figura 3.8). Este resultado es otra evidencia más a favor de la variación en la eficacia de la selección purificadora mediada por la tasa de recombinación y el efecto Hill-Robertson (Hill y Robertson 1966).

#### 4.2.5 EVOLUCIÓN RÁPIDA DEL X: MUTACIONES BENEFICIOSAS RECESIVAS

Este análisis ha permitido ayudar a esclarecer otras cuestiones controvertidas como la hipótesis de la evolución rápida del X (*faster-X effect*) que propone que el cromosoma X evoluciona a una tasa mayor que los autosomas debido a una mayor tasa de fijación de nuevas mutaciones adaptativas recesivas asumiendo un censo efectivo similar de machos y hembras (Avery 1984; Charlesworth *et al.* 1987; Betancourt *et al.* 2004, consúltese Meisel y Connallon 2013 para una revisión reciente).

Estudios previos que trataban de probar esta hipótesis en *Drosophila* habían llegado a conclusiones contradictorias (Thornton *et al.* 2006; Connallon 2007; Singh *et al.* 2007; Baines *et al.* 2008). La heterogeneidad en las muestras de estos trabajos podría explicar dichas discrepancias. En este sentido, el trabajo de Singh *et al.* (2008) es el primero en utilizar un conjunto de datos genómico para resolver dicha controversia en 12 especies de *Drosophila*. Sus resultados sugieren una mayor eficacia de la selección positiva en el X en algunas especies, entre ellas *D. melanogaster*, pero no en otras. En nuestro trabajo, después de corregir para las diferencias en la tasa de entrecruzamiento y la densidad génica entre X y autosomas, hemos encontrado que la tasa de adaptación promedio es ~1.4 veces mayor en el cromosoma X que en el resto de autosomas juntos (véase tabla 3.4). La razón de la tasa de adaptación X/autosomas por clase funcional (ordenando de mayor a menor) es la siguiente: (1) región intergénica (X/autosomas ~ 1,47), (2) UTR (X/autosomas ~ 1,45), (3) intrones (X/autosomas ~ 1,43) y (4) sitios no-sinónimos (X/autosomas ~ 1,30) (véase tabla 3.4). Esta mayor tasa de adaptación en el X es fruto, probablemente, de la mayor fijación de nuevas mutaciones beneficiosas parcialmente recesivas. Además, de acuerdo con el trabajo teórico de Orr y Betancourt (2001), si la adaptación se produjese sobre variantes preexistentes deletéreas (en equilibrio mutación-selección), los autosomas deberían mostrar mayores tasas de evolución adaptativa que el cromosoma X independientemente de la dominancia de las mutaciones (véase

sección 1.2.3). Por lo tanto, nuestros resultados también sugieren que una fracción substancial de las fijaciones adaptativas proviene de nuevas mutaciones.

Estudios a escala genómica posteriores a Singh *et al.* (2008) también observaron que la tasa de adaptación es mayor en el X que en los autosomas para una misma tasa de recombinación efectiva en *D. melanogaster* (Mackay *et al.* 2012; Langley *et al.* 2012; y Campos *et al.* 2014). Sin embargo, estos estudios siguen sin controlar para las diferencias en la densidad génica. En el trabajo de Campos *et al.* (2014), el cual utiliza el mismo estimador de la tasa de evolución adaptativa que nosotros (este es  $\omega_A$ ), las diferencias X/autosomas para la tasa de sustitución de mutaciones no-sinónimas adaptativas pasa de 1,88 (sin corregir) a 1,44 (corrigiendo la tasa de recombinación entre X y autosomas y comparando regiones de elevada recombinación en ambos casos). En nuestro caso la razón X/autosomas para tasa de substituciones no-sinónimas adaptativas sin corregir es de 1,43 (véase tabla 3.2), corregido sólo para la tasa de recombinación es 1,32 y corregido tanto para la densidad génica como para la recombinación es 1,30 (véase tabla 3.4). Estas ligeras diferencias entre nuestros resultados y los de Campos *et al.* (2014) pueden tener dos explicaciones no mutuamente excluyentes: (1) En nuestra población norteamericana el censo efectivo para el cromosoma X es  $\frac{2}{3}$  el de los autosomas (véase más arriba), la eficacia de la selección es menor y, por tanto, aunque nuestro estimador tiene en cuenta cambios demográficos recientes, cabe la posibilidad que nuestras estimas de la adaptación estén ligeramente subestimadas en el cromosoma X. En cambio, esto no es un problema en el conjunto de datos de Campos *et al.* (2014) basado en una población africana donde el censo efectivo X y autosomas es muy similar. (2) Nosotros hemos descartado las posiciones dos veces degeneradas porque hemos utilizado un criterio basado en las posiciones físicas para definir los sitios sinónimos y no-sinónimos (sección 2.2.3). Campos *et al.* (2014) utilizaron un criterio mutacional para definir los sitios sinónimos y no-sinónimos y por tanto no descartaron las posiciones dos veces degeneradas. Si una proporción importante de las nuevas mutaciones no-sinónimas

adaptativas y parcialmente recesivas se producen en las posiciones dos veces degeneradas entonces podemos estar subestimando la razón X/autosomas para la tasa de evolución adaptativa. Desafortunadamente, en el trabajo de Campos *et al.* (2014) sólo utilizaron genes codificadores y no podemos comparar sus resultados con los nuestros en el ADN no-codificador con tal de diferenciar entre estas dos hipótesis. Alternativamente, podríamos reanalizar nuestros datos definiendo los sitios sinónimos y no-sinónimos siguiendo un criterio mutacional como Campos *et al.* (2014) y comparar nuestros resultados.

La mayor divergencia en el ADN no-codificador que en el codificador en el cromosoma X respecto a los autosomas ha sido reportada utilizando otro conjunto de datos genómico en *D. melanogaster* (Hu *et al.* 2013). Nuestros resultados sugieren que esta mayor divergencia se debe a una mayor tasa de adaptación en el ADN no-codificador (y no a la relajación de la selección purificadora en el X). Existe una tercera evidencia que sale reforzada tras nuestro análisis; la evolución rápida de la expresión génica en el cromosoma X en *Drosophila* (Kayserili *et al.* 2012; Llopart 2012; Meisel *et al.* 2012). Esto se ha observado también en mamíferos (Khaitovich *et al.* 2005; Brawand *et al.* 2011). Hay evidencias que sugieren que la mayor tasa de evolución en los niveles de expresión de genes ligados al cromosoma X son debidos a la acción de la selección positiva y no el resultado de la deriva genética (Meiklejohn *et al.* 2003; Kayserili *et al.* 2012; Meisel *et al.* 2012). Nuestros resultados apuntan a que los cambios adaptativos reguladores en *cis* podrían ser los responsables de la evolución rápida de los niveles de expresión en el X de *D. melanogaster*.

Es posible que al haber utilizado una población norteamericana hayamos magnificado la razón de la tasa de adaptación X/autosomas. Es decir, nuestra población norteamericana es probable que se haya tenido que adaptar a nuevos ambientes y esto puede haber magnificado la razón de la tasa de adaptación X/autosomas. Sospechamos que este hecho será poco relevante, primero porque los estudios de

Langley *et al.* (2012) y Campos *et al.* (2014) utilizando poblaciones africanas también encuentran una mayor tasa de adaptación en el X, y segundo porque nuestras estimas de divergencia no están polarizadas, es decir, hemos estimado el total de substituciones entre *D. melanogaster* y *D. yakuba*. Las substituciones adaptativas fijadas únicamente en las poblaciones no-africanas serán, por tanto, una fracción menor de todas las substituciones adaptativas que distinguen a *D. melanogaster* y *D. yakuba*.

En conclusión, la mayor adaptación en el X respecto a autosomas parece un hecho para todas las poblaciones estudiadas de *D. melanogaster* (africanas y no-africanas) y es una prueba a favor de la existencia de nuevas mutaciones adaptativas parcialmente recesivas (muchas de ellas parecen reguladoras en *cis*) y de la hipótesis de la evolución rápida del X en *D. melanogaster*. Sin embargo, este hecho no parece generalizable a todas las especies del género *Drosophila* (Singh *et al.* 2008), lo cual sugiere que o bien la DFE o bien otros parámetros importantes como el grado de dominancia fenotípica de las mutaciones son muy variables entre especies y su interacción genera distintos resultados en cada linaje. Finalmente, otra fuente de evidencias a favor de la variación del  $N_e$  en el genoma, no ya entre cromosomas sino dentro de los mismos brazos cromosómicos, proviene de estudios donde se relaciona la tasa de recombinación y la tasa de adaptación en *Drosophila* (Presgraves 2005; Betancourt *et al.* 2009; Arguello *et al.* 2010; Mackay *et al.* 2012; Campos *et al.* 2014; Castellano *et al.* 2016). Los resultados mostrados en la sección 3.3, donde se incorpora la densidad génica y la tasa de mutación a la relación entre la tasa de recombinación y la tasa de adaptación (Castellano *et al.* 2016), se discuten en detalle más adelante (sección 4.3).

#### 4.2.6 INVERSIONES Y COMPOSICIÓN GÉNICA DIFERENCIAL ENTRE BRAZOS CROMOSÓMICOS

En este estudio no sólo hemos detectado diferencias en la tasa de adaptación y en la *DFE* de nuevas mutaciones deletéreas entre autosomas y X. Dentro de los autosomas hemos observado que el brazo cromosómico 2L muestra una tasa de substituciones adaptativas dos veces mayor al resto de brazos autosómicos juntos. Esto es probable que sea debido a la mayor tasa de recombinación en este brazo autosómico respecto al resto de autosomas (Comeron *et al.* 2012; Barrón 2015). A su vez, el brazo cromosómico 3R muestra en conjunto escasas evidencias de adaptación. Respecto a las diferencias en la *DFE*, el brazo cromosómico 3R muestra más mutaciones efectivamente neutras que el brazo cromosómico 2L. Dos explicaciones no mutuamente excluyentes podrían explicar estos resultados: (1) reducción de la tasa de recombinación debida a la presencia de inversiones polimórficas (Langley *et al.* 2012) y (2) una composición génica o patrones de expresión distintos entre brazos cromosómicos (Vicoso y Charlesworth 2006; Hu *et al.* 2013). Nótese que estas hipótesis también servirían para explicar el efecto del X rápido (véase más arriba).

La mayor cantidad de inversiones polimórficas (y la mayor proporción de cromosoma afectado por inversiones) en el brazo cromosómico 3R respecto al resto de brazos autosómicos y el cromosoma X (véase figura 3 de Corbett-Detig y Hartl 2012) puede conducir a una menor frecuencia de entrecruzamientos en la región invertida y con ello a una reducción de la eficacia de la selección natural (Langley *et al.* 2012). Además, las inversiones polimórficas pueden magnificar el alcance de los arrastres selectivos positivos al suprimir la recombinación entre grandes distancias - reduciendo más si cabe la eficacia de la selección (Andolfatto 2001). Este hecho podría explicar la mayor fracción de nuevas mutaciones efectivamente neutras, la menor proporción de mutaciones fuertemente deletéreas y la menor tasa de evolución adaptativa que hemos encontrado en el brazo cromosómico 3R. Sin embargo, el origen de todas estas inversiones polimórficas es relativamente reciente y es muy probable que no hayan tenido tiempo suficiente para afectar al número de

substituciones adaptativas que separan a *D. melanogaster* y *D. yakuba* (Campos *et al.* 2014), pero sí a los niveles de polimorfismo. Pool *et al.* (2012) encontraron en una población francesa de *D. melanogaster* que los niveles de variación nucleotídica aumentan en promedio cerca de un 30% en el brazo cromosómico 3R, un 20% en el brazo cromosómico 3L y un 10% en el brazo cromosómico 2L al incluir cromosomas portadores de inversiones (es decir, los cromosomas invertidos aportan variabilidad). Es muy probable que algo similar suceda con nuestra población norteamericana, pues la inversión *In(3R)Mo* (una de las tres descritas en el brazo cromosómico 3R) está al 12% en nuestra muestra DGRP (Langley *et al.* 2012; Corbett-Detig y Hartl 2012). Para las poblaciones africanas ancestrales las inversiones (las cuales se cree se originaron en África) tienen, como es de esperar, un impacto muy leve sobre los niveles de variación nucleotídica (Pool *et al.* 2012; Corbett-Detig y Hartl 2012); esto indicaría una llegada reciente de cromosomas invertidos provenientes de África y su incremento en frecuencia como consecuencia de la acción de la selección natural en poblaciones no africanas (Pool *et al.* 2012; Corbett-Detig y Hartl 2012). Sería muy interesante conocer el impacto de las inversiones polimórficas sobre el espectro de frecuencias de mutaciones seleccionadas y neutras. De esta forma podríamos evaluar cuantitativamente el impacto de las inversiones sobre nuestras estimas de la adaptación y la *DFE* – *a priori* esperamos subestimar la tasa de adaptación y la eficacia de la selección purificadora en las regiones afectadas por la inversión. El impacto de las inversiones sobre la tasa de evolución adaptativa no termina con el efecto que las inversiones polimórficas tienen sobre los niveles de variación nucleotídica actuales. Las inversiones fijadas entre *D. melanogaster* y *D. yakuba* se encuentra mayoritariamente en el brazo cromosómico 3R (Lemeunier y Aulard 1992). Además, habrán existido inversiones que fueron polimórficas en algún momento y acabaron por extinguirse, todas estas también habrán contribuido a la reducción de la recombinación y la eficacia de la selección en algún momento. Evaluar el impacto de estos casos sobre nuestras estimas es, sin embargo, muy complicado o virtualmente imposible.

Otro potencial factor que podría explicar las diferencias en la DFE y la tasa de adaptación entre cromosomas es la composición génica de estos. Hu *et al.* (2013) se preguntaron si los términos GO (*gene ontology*) asociados a genes codificadores podrían explicar las diferencias X/autosomas (y entre autosomas) en los niveles de divergencia específica del linaje de *D. melanogaster* (y *D. simulans*). Realizaron un ANOVA donde los términos GO eran un factor más y encontraron que entre autosomas los términos GO no contribuían a explicar la variación en los niveles de divergencia en *D. melanogaster* (ni *D. simulans*). En cambio, la variación en los niveles de divergencia no-sinónima entre X y autosomas sí se podían explicar enteramente a partir de los términos GO en *D. melanogaster* (y en *D. simulans*). Este hecho es una evidencia en contra de la mayor tasa de evolución en el X debida a la fijación de mutaciones beneficiosas parcialmente recesivas en los sitios no-sinónimos pues sugiere que el cromosoma X evoluciona más rápido debido a su composición génica. Curiosamente, los términos GO no contribuyen a explicar la variación en los niveles de divergencia en sitios no-codificadores entre X/autosomas (ni entre autosomas) en *D. melanogaster* (ni *D. simulans*) (Hu *et al.* 2013), evidenciando que las mutaciones beneficiosas parcialmente recesivas en sitios no-codificadores sí podrían ser la explicación a la mayor divergencia encontrada en estos sitios para el cromosoma X. A su vez, los genes de expresión sesgada en machos se sabe están subrepresentados en el cromosoma X y se ha descrito que exhiben elevadas tasas de evolución adaptativa en *Drosophila* (Sturgill *et al.* 2007; Zhang *et al.* 2007). Hu *et al.* (2013) realizaron un ANOVA con tal de saber si los genes de expresión sesgada en machos contribuían a explicar la variación en los niveles de divergencia para secuencias codificadoras y no-codificadoras entre cromosomas. No encontraron evidencias a favor de esta hipótesis.

En conclusión, las diferencias entre autosomas en lo que respecta a las estimas de la DFE y la tasa de adaptación podrían deberse a la mayor prevalencia de inversiones polimórficas y fijadas en algunos brazos autosómicos (especialmente en el brazo

cromosómico 3R) y a la mayor tasa de recombinación en el brazo 2L (Comeron *et al.* 2012; Barrón 2015). Ni las diferencias en la composición génica ni los patrones de expresión parecen fuertes candidatos para explicar las diferencias en la DFE y la tasa de adaptación entre autosomas (basándonos en el trabajo de Hu *et al.* 2013). Sin embargo, las diferencias en la divergencia no-sinónima sí se explican enteramente por diferencias en la composición génica entre X y autosomas (Hu *et al.* 2013). Por lo tanto, cabe la posibilidad que la mayor tasa de evolución adaptiva en sitios no-sinónimos en el cromosoma X pueda ser debida también a las diferencias en la composición génica y no a una mayor fijación de nuevas mutaciones parcialmente beneficiosas en esta clase de sitios como apuntábamos anteriormente. En futuros análisis deberíamos tener en cuenta la composición génica y reevaluar nuestras conclusiones.

## 4.3 EFECTO HILL-ROBERTSON SOBRE LA TASA DE EVOLUCIÓN ADAPTATIVA EN PROTEÍNAS DE *D. melanogaster*

Se ha mostrado como la tasa de evolución adaptativa en proteínas está positivamente correlacionada con la tasa de recombinación y la tasa de mutación, pero negativamente correlacionada con la densidad génica en *D. melanogaster*. Estas correlaciones no son debidas a un enriquecimiento de genes del sistema inmune o de expresión sesgada en machos en regiones de baja densidad génica o alta recombinación y mutación, o debidas a la selección en el uso de codones. En su lugar, es probable que la tasa de evolución adaptativa esté correlacionada positivamente con la tasa de recombinación y negativamente con la densidad génica a raíz de la interferencia de Hill-Robertson (iHR). A su vez, es probable que la tasa de evolución adaptativa esté correlacionada positivamente con la tasa de mutación debido a que los genes con elevadas tasas de mutación es más probable que generen la diversidad genética necesaria para la adaptación. No obstante, la relación positiva entre la tasa de adaptación y la tasa de mutación se desvanece para aquellos genes localizados en regiones de baja recombinación o de alta densidad génica, este hecho confirma que la iHR es más prevalente cuando el número de mutaciones seleccionadas es alto y/o la distancia genética entre ellas es pequeña.

En este trabajo se ha cuantificado por primera vez en una especie el impacto global de la iHR sobre un genoma. Aproximadamente el 27% de todas las mutaciones adaptativas, que se deberían de haber fijado con libre recombinación entre sitios, se han extinguido debido a la iHR. Esta fracción de mutaciones beneficiosas perdidas depende de la tasa de mutación y de la densidad génica; genes con elevadas tasas de mutación localizados en regiones ricas en genes pierden una mayor proporción de sus substituciones adaptativas (~60%) que los genes con bajas tasas de mutación localizados en regiones pobres en genes (~17%).

#### 4.3.1 ¿ES LA TASA DE ADAPTACIÓN PROTEICA INDEPENDIENTE DE LA TASA DE MUTACIÓN?

Las mutaciones ligeramente beneficiosas ( $1 < N_e s < 10$ ) requieren más tiempo para fijarse y son más susceptibles a extinguirse debido a la disminución del  $N_e$  promovida por la iHR y/o cuellos de botella demográficos que las mutaciones fuertemente beneficiosas ( $N_e s > 10$ ), las cuales se fijarán más fácilmente. Karasov *et al.* (2010) sugieren que las etapas de *boom* demográfico podrían ser las más comunes en la historia demográfica de *D. melanogaster* (véase cuadro 4). La cantidad de individuos en la población en estas etapas de *boom* podría ser de tal magnitud que posibilitaría tasas de mutación poblacionales por sitio mayores a la unidad, es decir, que en una misma generación (o en pocas generaciones de diferencia) el mismo tipo de mutación podría ocurrir de manera independiente en distintos individuos (y *backgrounds* genéticos). Si esto es cierto y además la mayoría de mutaciones adaptativas están fuertemente seleccionadas entonces esperaríamos una falta de correlación entre la tasa de adaptación y la tasa de mutación, pues esta última no sería limitante. Si por el contrario la mayoría de las nuevas mutaciones adaptativas están débilmente seleccionadas (como sugiere Scheneider *et al.* 2011) entonces sí esperamos encontrar, tal y como observamos, una correlación positiva entre mutación y adaptación, pues la iHR y/o los cuellos demográficos eliminarían muchas variantes beneficiosas de la población cada generación y habrá períodos con escasas mutaciones beneficiosas. En otras palabras, el hecho de haber encontrado una relación positiva entre adaptación y mutación sugiere que una proporción notable de las nuevas mutaciones beneficiosas están débilmente seleccionadas en *Drosophila*. Este resultado no indica, no obstante, que la adaptación ocurra sobre nuevas variantes, sólo señala que la adaptación está limitada por la entrada de nuevas mutaciones.

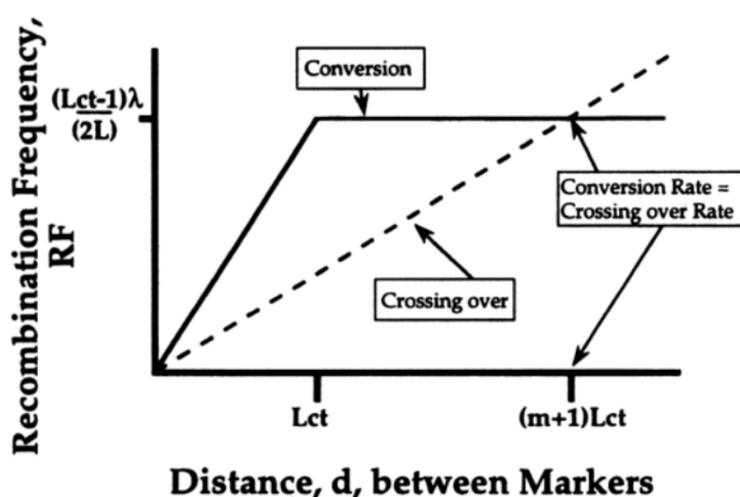
#### 4.3.2 iHR Y CONSTANCIA DE LA DFE A LO LARGO DEL GENOMA

Bajo el modelo más sencillo de iHR, la DFE (de nuevas mutaciones deletéreas y adaptativas) permanece estable a lo largo del genoma, y lo único que varía es la tasa

de recombinación, la densidad génica y la tasa de mutación. Sin embargo, en la práctica esto es poco probable que suceda, pues la iHR reduce el  $N_e$  localmente afectando nuestras estimas de la *DFE* subyacente. La interferencia de Hill-Robertson reduce el  $N_e$  de una región, pero no afecta a todas las mutaciones seleccionadas por igual; los arrastres positivos reducen más el  $N_e$  para mutaciones fuertemente deletéreas que para mutaciones ligeramente deletéreas (Messer y Petrov 2013). Por lo tanto, esperamos que la iHR afecte a los parámetros que definen la distribución gamma que hemos asumido para modelar la *DFE* de nuevas mutaciones deletéreas de una forma compleja. Tanto nuestras estimas del parámetro de forma ( $\beta$ ) como del efecto promedio de las nuevas mutaciones deletéreas ( $\overline{N_e s}$ ) se verán afectadas por la iHR. Pero ¿cómo esperamos que la iHR afecte a  $\beta$  y  $\overline{N_e s}$ ? A medida que aumenta la intensidad de la iHR esperaríamos una reducción en  $\overline{N_e s}$ , lo cual conduce a un aumento de las mutaciones que son efectivamente neutras y una reducción de las mutaciones fuertemente deletéreas, lo que sucede con la fracción de mutaciones ligeramente deletéreas dependerá de la forma de la distribución la cual viene definida por el parámetro  $\beta$ . Es decir, *a priori* es difícil predecir qué pasará con el parámetro  $\beta$  bajo un  $N_e$  fluctuante. Como esperamos, hemos encontrado que nuestras estimas del  $\overline{N_e s}$  están positivamente correlacionadas con la tasa de recombinación y negativamente correlacionadas con la tasa de mutación y la densidad génica ( $\overline{N_e s}$  vs tasa de recombinación  $\rho_s = 0,48, P < 0,001$ ;  $\overline{N_e s}$  vs tasa de mutación  $\rho_s = -0,36, P < 0,05$ ;  $\overline{N_e s}$  vs densidad génica  $\rho_s = -0,23, P = 0,12$ ). En cualquier caso, y lo que es más importante, nuestro método para estimar la tasa de evolución adaptativa tiene en cuenta las diferencias intrínsecas y extrínsecas (debidas a la iHR) en la *DFE* entre grupos de genes. No esperamos, por tanto, que la variación en la *DFE* afecte a nuestras estimas de la tasa de adaptación ni a nuestras conclusiones.

#### 4.3.3 ESCASA IMPORTANCIA DE LA CONVERSIÓN GÉNICA PARA LA iHR

Cada suceso de recombinación ( $c$ ) puede resolverse como conversión génica asociado a entrecruzamiento o como sucesos de conversión génica sin entrecruzamiento (Mortimer y Fogel 1974, Foss *et al.* 1993, Navarro *et al.* 1996). Berry y Barbadilla (2000) argumentan sobre la importancia relativa de estos dos procesos básicos de recombinación a el nivel de la secuencia de ADN (figura 4.2).



**FIGURA 4.2** Gráfico que relaciona la distancia entre dos marcadores y su probabilidad que recombinen por conversión génica o por entrecruzamiento.  $L$ : Longitud del cromosoma,  $Lct$ : Longitud del tracto de conversión,  $m$ : Razón de los sucesos de conversión respecto a los sucesos de entrecruzamiento,  $\lambda$ : Tasa de conversión por generación. [Figura tomada de Berry y Barbadilla (2000)].

Al considerar las variables longitud promedio de un tracto de conversión ( $Lct$ ) y la razón entre la tasa de conversión génica y la de entrecruzamiento ( $m+1$ ), los autores distinguen tres regiones que caracterizan la importancia relativa de la conversión vs el entrecruzamiento como causas de la recombinación entre marcadores: (1) para distancias cortas entre dos marcadores,  $d$  (distancia entre marcadores)  $< Lct$ , la conversión génica es más importante que el entrecruzamiento como fuente de recombinación. (2) A distancias en las que  $d$  es mayor que  $Lct$  pero menor que  $(m+1)Lct$ , la conversión génica sigue generando más recombinación que el

entrecruzamiento, aunque el efecto del entrecruzamiento se aproxima al de la conversión linealmente con la distancia. Por último, (3) a distancias largas, mayores que  $(m+1)Lct$ , el entrecruzamiento se convierte en el proceso dominante de recombinación.

La razón de la frecuencia de los sucesos de conversión génica sin entrecruzamiento respecto a la frecuencia de los sucesos que incluyen el entrecruzamiento se ha estimado experimentalmente que es  $\sim 4$  (Hilliker y Chovnick 1981; Hilliker *et al.* 1991; Foss *et al.* 1993; Comeron *et al.* 2012). Es decir, que en *Drosophila* de cada 5 sucesos de conversión génica ocurre un suceso de entrecruzamiento. Además, Comeron *et al.* (2012) calcula que la longitud promedio del tracto de conversión del genoma es de  $\sim 518$  pb. Teniendo en cuenta estos valores observados, podemos suponer que la conversión génica es 5 veces más importante que el entrecruzamiento en regiones menores que 518 pb. Por tanto, la distancia crítica a partir de la cual el entrecruzamiento empieza a ser la fuerza principal generadora de recombinación sería:  $518 \times 5 = 2590$  pb. Es decir, para distancias menores a  $\sim 2,5$  kb la conversión génica sería la causa principal de recombinación, para distancias mayores la tasa de entrecruzamiento define el alcance de la iHR.

En este trabajo hemos utilizado únicamente tasas de entrecruzamiento y hemos excluido eventos de conversión génica (estimas provenientes de Comeron *et al.* 2012). Al hacer esto hemos confirmado empíricamente lo que la teoría sugiere: la conversión génica es poco eficiente reduciendo la iHR, ya que, una fracción substancial de las substituciones adaptativas dejan de fijarse en regiones del genoma con escasos eventos de entrecruzamiento (como regiones cercanas al centrómero) pero que disfrutan de niveles notables de conversión génica – la frecuencia de la conversión génica varía muy poco a lo largo del genoma de *D. melanogaster* (véase cuadro 5 en sección 2.4) (Comeron *et al.* 2012).

#### **4.3.4 POSIBLE IMPACTO DE LA INTERFERENCIA DE HILL-ROBERTSON EN LA ADAPTACIÓN HUMANA**

Una pregunta que interesante es hasta qué punto la iHR afectará a las tasas de evolución adaptativa en otras especies, en concreto en la especie humana. La intensidad de la iHR depende de la tasa de mutación en sitios putativamente seleccionados, la *DFE* y la tasa de recombinación; cuanto mayor sea la densidad de mutaciones seleccionadas por unidad de recombinación, y cuanto más intensamente seleccionadas estén las mutaciones, mayor será la iHR sobre mutaciones débilmente seleccionadas. ¿Cuán probable es que la iHR sea una fuerza relevante en la especie humana? Los humanos tenemos una tasa genómica de mutaciones deletéreas de 2,1 (Lesecque *et al.* 2012), aproximadamente el doble que *Drosophila* con 1,2 mutaciones deletéreas por genoma (Haag-Liautard *et al.* 2007). Aunque el genoma humano es unas 20 veces mayor que el genoma de *Drosophila*, el desequilibrio de ligamiento decae unas 500 veces más lentamente en humanos que en *Drosophila*. Estos datos sugieren que la iHR (al menos para mutaciones deletéreas) podría ser más importante en humanos que en *Drosophila*. No obstante, nótese que el genoma humano disfruta de más cromosomas que *Drosophila*, la segregación entre mutaciones localizadas en distintos cromosomas es por definición libre. Es necesario evaluar esta hipótesis mediante análisis, y esto es complicado debido a que las estimas de la tasa de evolución adaptativa en la especie humana parecen ser muy bajas (Boyko *et al.* 2008; Eyre-Walker y Keightley 2009; Gossmann *et al.* 2012); cualquier análisis de los factores que afectan a la tasa de evolución adaptativa en humanos se verá dificultado por este hecho. Sin embargo, una mayor intensidad de la iHR en humanos podría explicar porque nuestra especie parece haber sufrido pocos cambios adaptativos a nivel de secuencia en comparación con *Drosophila* (Gossmann *et al.* 2012). Además, el efecto Hill-Robertson depende también de la *DFE* (tanto de mutaciones adaptativas como deletéreas) y sobre la *DFE* disponemos de muy poca información en estas y otras especies. Realizar en otras especies el análisis llevado aquí en *Drosophila* será de gran utilidad para entender el impacto de la iHR en la adaptación en términos más generales.

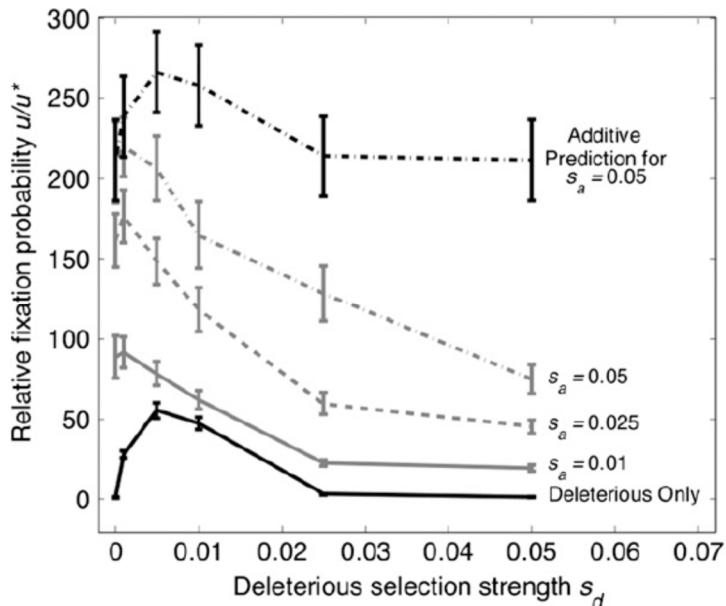
#### **4.3.5 ¿NUEVO MÉTODO PARA LA INFERENCIA DE LA DFE DE NUEVAS MUTACIONES BENEFICIOSAS?**

Como las mutaciones más débilmente seleccionadas son las más afectadas por la iHR (McVean y Charlesworth 2000; Comeron y Kreitman 2002; Comeron *et al.* 2008), la pérdida de substituciones adaptativas debido a la iHR nos informa sobre la fuerza de la selección que actúa sobre algunas mutaciones beneficiosas. Para estimar los parámetros relevantes para la adaptación, estos son la fracción de nuevas mutaciones que son adaptativas ( $p_a$ ) y el efecto promedio de las nuevas mutaciones adaptativas ( $N_{eS_a}$ ), podríamos utilizar la fracción de substituciones adaptativas no fijadas debido a la iHR. Para hacer esto deberíamos simular la *DFE* de mutaciones beneficiosas y deletéreas ante distintas tasas de recombinación y estudiar qué mutaciones adaptativas son las más susceptibles a la iHR. Esta información junto con: (1) el decaimiento de la diversidad genética neutra alrededor de substituciones en sitios putativamente seleccionados (Sella *et al.* 2009; Lee *et al.* 2014), (2) el espectro de frecuencias de alelos derivados (Schneider *et al.* 2011) y (3) el exceso de variantes a alta frecuencia tras un arrastre selectivo positivo (Fay y Wu 2000), podría hacer posible conocer más detalles de la *DFE* de nuevas mutaciones adaptativas de lo que hasta ahora creíamos.

#### **4.3.6 SELECCIÓN SOBRE MODIFICADORES DE LA TASA DE RECOMBINACIÓN**

El hecho que tantas substituciones adaptativas se hayan perdido debido a la iHR plantea la cuestión: ¿Por qué *D. melanogaster* no tiene mayores tasas de recombinación? Sobre todo en regiones de baja o nula recombinación como los centrómeros. Simulaciones por ordenador muestran que en poblaciones finitas la probabilidad de fijación de modificadores de la recombinación (esto son *loci* que aumentan la tasa de recombinación) en presencia de otras mutaciones seleccionadas (beneficiosas y deletéreas, o sólo deletéreas, o sólo beneficiosas) aumenta respecto a lo esperado para una mutación que no afecta a la recombinación entre *loci* (Hartfield *et al.* 2010; Keightley y Otto 2006; Iles *et al.* 2003). Siempre que haya desequilibrio de

ligamiento entre alelos seleccionados, habrá una selección indirecta para favorecer la fijación del modificador de la recombinación e incrementar los eventos de recombinación en la población. La intensidad de dicha selección (indirecta) sobre el modificador de la recombinación dependerá de una serie de parámetros y/o condiciones (figura 4.3).



**FIGURA 4.3** Probabilidad relativa de fijación de un modificador de la selección para  $N = 25,000$  individuos haploides en función de la fuerza de la selección purificadora ( $s_d$ ).  $u/u^*$  representa la probabilidad de fijación observada ( $u$ ) en relación a una nueva mutación neutra ( $u^*$ ). Las mutaciones son sólo deletéreas o una mezcla de deletéreas y beneficiosas con coeficientes  $s_a$ . [Figura tomada de Hartfield *et al.* (2010)].

La mayor ventaja para los modificadores de la recombinación se produce en presencia de mutaciones deletéreas débilmente seleccionadas y mutaciones beneficiosas fuertemente seleccionadas (figura 4.3). La ventaja que supone la recombinación en este escenario es doble, por un lado, elimina más eficientemente las mutaciones deletéreas y por otra fija más mutaciones beneficiosas. Incluso en el peor escenario, esto es sólo en presencia de mutaciones fuertemente deletéreas (véase línea continua de la figura 4.3), la probabilidad de fijación del modificador es mayor a lo

esperado para una nueva mutación neutra,  $u/u^* \sim 1,32$  (Hartfield *et al.* 2010). Quizás los genes que se encuentran en las regiones con escasa recombinación en *D. melanogaster* (como los centrómeros) disponen de una DFE tal que los modificadores de la recombinación encuentran escasas oportunidades para establecerse (es decir, todas las mutaciones que allí ocurren son fuertemente deletéreas y no hay mutaciones beneficiosas). Esta hipótesis parece poco parsimoniosa pues implicaría una ordenación espacial de los genes en los cromosomas de acuerdo a su DFE.

La supresión de los entrecruzamientos cerca de los centrómeros en autosomas no es una propiedad exclusiva de *D. melanogaster*, otras muchas especies muestran este patrón (Jones 1987). Nótese que la mayoría de elementos transponibles en *D. melanogaster* se encuentran en estas regiones de baja recombinación centroméricas. Se ha sugerido que esto evitaría eventos de recombinación ectópica entre elementos y la consecuente formación de cromosomas aberrantes muy deletéreos (Montgomery *et al.* 1987; Langley *et al.* 1988; Charlesworth *et al.* 1992). De hecho, se ha descrito que algunas especies del género *Drosophila* que no muestran esta supresión de los entrecruzamientos cerca de los centrómeros, como *D. mauritania* (True *et al.* 1996), tienen una menor actividad y cantidad de elementos transponibles en su genoma (Kofler y Schlötterer 2015). Aunque bajo esta segunda hipótesis se pierde una gran cantidad de substituciones beneficiosas (un 27% del total según nuestros cálculos), esto supondría un mal menor en comparación con los efectos negativos sobre la *fitness* que acarrea la generación de aberraciones cromosómicas. Otra cuestión interesante íntimamente relacionada con esta es por qué los genes no escapan de regiones de baja recombinación relocalizándose en regiones de alta recombinación. Esto es probablemente porque sólo pueden fijarse en regiones de alta recombinación si se translocan a un *locus* donde ya hay una mutación beneficiosa propagándose en la población, o donde la mutación beneficiosa empieza a propagarse poco después de la translocación. La co-ocurrencia de ambos fenómenos será algo muy poco probable.

# CONCLUSIONS

---

## NEW POINT ESTIMATORS OF PURIFYING SELECTION

1. We proposed two new estimators ( $d_n$  and  $b$ ) of the action of recent purifying selection based on nucleotide polymorphism which are qualitative estimators of the underlying DFE for new deleterious mutations.
2.  $d_n$  and  $b$  estimates are highly correlated to two previous maximum likelihood estimators of the fraction of strongly and weakly selected new deleterious mutations, respectively.
3.  $d_n$  is robust to non-equilibrium conditions while  $b$  estimates are biased under a recent population bottleneck.
4. Both estimators are reliable when tens or hundreds of segregating sites are available.

## DFE AND ADAPTIVE EVOLUTION ALONG THE CODING AND NON-CODING GENOME OF *D. melanogaster*

5. Both our new estimators and previous estimators of the DFE show that purifying selection is pervasive along the coding and non-coding genome of *D. melanogaster*. We estimate that ~ 60% of the *D. melanogaster* genome is under recent purifying selection (or functional).
6. The majority of slightly deleterious mutations occur on non-coding sequences. However, nonsynonymous mutations explain most of the variance in fitness between individuals because the average strength of selection is more intense for coding than non-coding mutations.
7. The rate of adaptive substitutions in non-coding sequences is on average two fold the estimate in coding regions. Interestingly, no differences in the rate of adaptive evolution between non-coding class sites (UTR, introns and intergenic regions) are

- found. Regulatory mutations might dominate the adaptive change.
8. The fraction of slightly deleterious nonsynonymous mutations is lower in the X chromosome than in the autosomes while the rate of adaptive evolution is higher in the X chromosome than in the autosomes for both coding and non-coding mutations. This can be accounted for the presence of partially recessive new deleterious and beneficial mutations, respectively.
  9. There are some evidences which indicate that the efficacy of selection (negative and positive) is lower in the chromosome arm 3R relative to the rest of chromosome arms. The presence of more polymorphic (and fixed) large inversions in this autosome arm compared to the rest of chromosome arms could explain this result.

#### HILL-ROBERTSON INTERFERENCE (HRI) AND ADAPTIVE PROTEIN EVOLUTION IN *D. melanogaster*

10. We find that the rate of adaptive amino acid substitution is positively correlated to both recombination rate and an estimate of the mutation rate, while it is negatively correlated to gene density. These genomic variables are important determinants of the HRI.
11. The rate of protein adaptation seems to be limited by the supply of new mutations in *D. melanogaster*.
12. We estimate that on average at least ~27% of all advantageous substitutions have been lost because of HRI and that this quantity depends on gene's mutation rate and the gene density where the gene is located: genes with low mutation rates embedded in gene poor regions lose ~17% of their adaptive substitutions while genes with high mutation rates embedded in gene rich regions lose ~60%.

## BIBLIOGRAFÍA

- 1000 Genomes Project Consortium. *A global reference for human genetic variation*. Nature 467: 1061–1073 (2010)
- 1000 Genomes Project Consortium. *An integrated map of genetic variation from 1,092 human genomes*. Nature 491: 56–65 (2012)
- 1000 Genomes Project Consortium. *A global reference for human genetic variation*. Nature 526: 68–74 (2015)
- Adams MD, et al. *The genome sequence of Drosophila melanogaster*. Science 287(5461): 2185–2195 (2000)
- Akashi H. *Synonymous codon usage in Drosophila melanogaster: natural selection and translational accuracy*. Genetics 136: 927–935 (1994)
- Akashi H. *Inferring weak selection from patterns of polymorphism and divergence at "silent" sites in Drosophila DNA*. Genetics 139: 1067–1076 (1995)
- Akashi H and Schaeffer S. *Natural selection and the frequency distributions of "silent" DNA polymorphism in Drosophila*. Genetics 146: 295–307 (1997)
- Akashi H. *Inferring the fitness effects of DNA mutations from polymorphism and divergence data: statistical power to detect directional selection under stationarity and free recombination*. Genetics 151: 221–238 (1999)
- Akey JM, et al. *Population history and natural selection shape patterns of genetic variation in 132 genes*. PLoS Biol. 2: 1591–1599 (2004)
- Akey JM. *Constructing genomic maps of positive selection in humans: where do we go from here?* Genome Res. 19: 711–722 (2009)
- Alexander RP, et al. *Annotating non-coding regions of the genome*. Nat Rev Genet. 11: 559–71 (2010)
- Allendorf FW. *Genetic drift and the loss of alleles versus heterozygosity*. Zoo Biol. 5: 181–190 (1986)
- Allison AC. *Protection afforded by sickle-cell trait against subtertian malarial infection*. Br Med J. 1(4857): 290–294 (1954)
- Altshuler D, et al. *A map of human genome variation from population-scale sequencing*. Nature 467: 1061–1073 (2010)
- Andolfatto P. *Contrasting patterns of X-linked and autosomal nucleotide variation in Drosophila melanogaster and Drosophila simulans*. Mol Biol Evol. 18(3): 279–290 (2001)
- Andolfatto P. *Adaptive evolution of non-coding DNA in Drosophila*. Nature 437: 1149–1152 (2005)
- Andolfatto P. *Hitchhiking effects of recurrent beneficial amino acid substitutions in the Drosophila melanogaster genome*. Genome Res. 17: 1755–1762 (2007)
- Andolfatto P, et al. *Effective population size and the efficacy of selection on the X chromosomes of two closely related Drosophila species*. Genome Biol. Evol. 3, 114–28 (2011)
- Antezana MA and Kreitman M. *The nonrandom location of synonymous codons suggests that reading frame-independent forces have patterned codon preferences*. J Mol Evol. 49: 36–43 (1999)
- Ashburner M and Bergman CM. *Drosophila melanogaster: a case study of a model genomic sequence and its consequences*. Genome Res. 15(12): 1661–1667 (2005)

- Ashburner M, et al. *Drosophila: a laboratory hand-book*. 2nd ed. New York: Cold Spring Harbor Laboratory Press (2005)
- Assaf ZJ, et al. *Obstruction of adaptation in diploids by recessive, strongly deleterious alleles*. Proc Natl Acad Sci USA 112: E2658–2666 (2015)
- Aquadro CF, et al. *Selection, recombination, and DNA polymorphism in Drosophila*. In B. Golding, (ed.) *Non-neutral evolution*. Chapman and Hall, New York, pp. 46–56 (1994)
- Arguello JR, et al. *Recombination yet inefficient selection along the Drosophila melanogaster subgroup's fourth chromosome*. Mol Biol Evol. 27: 848–861 (2010)
- Avery PJ. *The population genetics of haploid-diploids and X-linked genes*. Genet Res. 44: 321–341 (1984)
- Avise JC and Lansman RA. *Polymorphism of mitochondrial DNA in populations of higher animals*. In Nei M. and Koehn RK (ed.) *Evolution of Gene and Proteins*, Sinauer, Sunderland, Mass., pp. 147–164 (1983)
- Ayala FJ, et al. *Genetic Variation in Natural Populations of Five Drosophila Species and the Hypothesis of the Selective Neutrality of Protein Polymorphisms*. Genetics 77(2): 343–384 (1974)
- Bachtrog D and Andolfatto P. *Selection, recombination and demographic history in Drosophila miranda*. Genetics. 174: 2045–2059 (2006)
- Bachtrog D. *Similar rates of protein adaptation in Drosophila miranda and D. melanogaster, two species with different current effective population sizes*. BMC Evol Biol. 8: 334 (2008)
- Baines JF, et al. *Pleiotropic effect of disrupting a conserved sequence involved in a long-range compensatory interaction in the Drosophila Adh gene*. Genetics 166: 237–242 (2004)
- Baines JF, et al. *Effects of X-linkage and sex-biased gene expression on the rate of adaptive protein evolution in Drosophila*. Mol Biol Evol. 25 (8): 1639–1650 (2008)
- Banks SC, et al. *How does ecological disturbance influence genetic diversity?* Trends Ecol Evol. 28: 670–679 (2013)
- Barrett RDH and Schlüter D. *Adaptation from standing genetic variation*. Trends Ecol Evol. 23: 38–44 (2008)
- Barrón MG. *Nucleotide variation patterns and linked selection blocks mapping along the Drosophila melanogaster genome* [PhD thesis]. [Barcelona (Spain)]: Universitat Autònoma de Barcelona. Available from: <http://hdl.handle.net/10803/310424> (2015)
- Barton N. *The divergence of a polygenic system subject to stabilizing selection, mutation and drift*. Genet Res. 54: 59–77 (1989)
- Barton NH. *Linkage and the limits to natural selection*. Genetics 140: 821–841 (1995)
- Barton NH. *Genetic hitchhiking*. Philos Trans R Soc Lond B Biol Sci. 355: 1553–1562 (2000)
- Barton NH. *Genetic linkage and natural selection*. Philos Trans R Soc Lond B Biol Sci. 365: 2559–2569 (2010)
- Baudry E and Depaulis F. *Effect of misoriented sites on neutrality tests with outgroup*. Genetics 165: 1619–1622 (2003)
- Baudry E, et al. *Non-African populations of Drosophila melanogaster have a unique origin*. Mol Biol Evol. 21 (8): 1482–1491 (2004)

- Bauer VL and Aquadro CF. *Rates of DNA sequence evolution are not sex-biased in Drosophila melanogaster and D. simulans*. Mol Biol Evol. 14: 1252–1257 (1997)
- Bazin E, et al. *Population size does not influence mitochondrial genetic diversity in animals*. Science 312: 570–572 (2006)
- Beaumont MA, et al. *Approximate Bayesian computation in population genetics*. Genetics 162(4): 2025–2035 (2002)
- Becquet C and Przeworski M. *A new approach to estimate parameters of speciation models with application to apes*. Genome Res. 17(10): 1505–1519 (2007)
- Begun DJ and Aquadro CF. *Levels of naturally occurring DNA polymorphism correlate with recombination rates in Drosophila melanogaster*. Nature 356: 519–520 (1992)
- Begun DJ and Aquadro CF. *African and North American populations of Drosophila melanogaster are very different at the DNA level*. Nature 365: 548–550 (1993)
- Begun DJ. *Population genetics of silent and replacement variation in Drosophila simulans and D. melanogaster: X/autosome differences?* Mol Biol Evol. 13: 1405–1407 (1996)
- Begun DJ and Whitley P. *Reduced X-linked nucleotide polymorphism in Drosophila simulans*. Proc Natl Acad Sci. USA 97: 5960–5965 (2000)
- Begun DJ, et al. *Population genomics: Whole-genome analysis of polymorphism and divergence in Drosophila simulans*. PLoS Biol. 5: e310 (2007)
- Berger J, et al. *Genetic mapping with SNP markers in Drosophila*. Nature Genet. 29(4): 475–481 (2001)
- Bergland AO, et al. *Genomic evidence of rapid and stable adaptive oscillations over seasonal time scales in Drosophila*. PLoS Genet. 10: e1004775 (2014)
- Bergman CM and Kreitman M. *Analysis of conserved noncoding DNA in Drosophila reveals similar constraints in intergenic and intronic sequences*. Genome Res. 11: 1335–1345 (2001)
- Bergman CM, et al. *Assessing the impact of comparative genomic sequence data on the functional annotation of the Drosophila genome*. Genome Biol. 3: research0086.1–0086.20 (2002)
- Bergman CM, et al. *Drosophila DNase I footprint database: a systematic genome annotation of transcription factor binding sites in the fruitfly, Drosophila melanogaster*. Bioinformatics 21(8): 1747–1749 (2005)
- Berry A and Barbadilla A. *Gene conversion is a major determinant of genetic diversity at the DNA level*. In Evolutionary Genetics: From Molecules to Morphology, (ed.) R.S. Singh and C.B. Krimbas. Cambridge University Press, USA. pp. 102–123 (2000)
- Betancourt AJ, et al. *A pseudohitchhiking model of x vs. autosomal diversity*. Genetics. 168: 2261–2269 (2004)
- Betancourt AJ, et al. *Reduced effectiveness of selection caused by a lack of recombination*. Curr Biol. 19: 655–660 (2009)
- Bierne N and Eyre-Walker A. *The problem of counting sites in the estimation of the synonymous and nonsynonymous substitution rates: implications for the correlation between the synonymous substitution rate and codon usage bias*. Genetics 165: 1587–97 (2003)

- Bierne N and Eyre-Walker A. *The genomic rate of adaptive amino acid substitution in Drosophila*. Mol Biol Evol 21: 1350–1360 (2004)
- Bierne N and Eyre-Walker A. *Variation in synonymous codon use and DNA polymorphism within the Drosophila genome*. J Evol Biol. 19: 1–11 (2006)
- Binns D, et al. *QuickGO: a web-based tool for Gene Ontology searching*. Bioinformatics 25: 3045–3046 (2009)
- Birky Jr CW and Walsh JB. *Effects of linkage on rates of molecular evolution*. Proc Natl Acad Sci. USA 85: 6414–6418 (1988)
- Blow MJ, et al. *ChIP-Seq identification of weakly conserved heart enhancers*. Nat Genet. 42(9): 806–810 (2010)
- Bonhomme M, et al. *Detecting selection in population trees: the Lewontin and Krakauer test extended*. Genetics 186(1): 241–262 (2010)
- Boyko AR, et al. *Assessing the evolutionary impact of amino acid mutations in the human genome*. PLoS Genet. 4: e1000083 (2008)
- Brawand D, et al. *The evolution of gene expression levels in mammalian organs*. Nature 478: 343–348 (2011)
- Bromham L, et al. *Determinants of rate variation in mammalian DNA sequence evolution*. J Mol Evol. 43: 610–621 (1996)
- Burbano HA, et al. *Analysis of human accelerated DNA regions using archaic hominin genomes*. PLoS One 7(3): e32877 (2012)
- Bustamante CD, et al. *The cost of inbreeding in Arabidopsis*. Nature 416: 531–534 (2002)
- Bustamante CD, et al. *Natural selection on protein-coding genes in the human genome*. Nature 437(7062): 1153–1157 (2005)
- Caballero A. *Developments in the prediction of effective population size*. Heredity 73: 657–679 (1994)
- Cai Z, et al. *Identification of regions of positive selection using shared genomic segment analysis*. Eur J Hum Genet. 19(6): 667–671 (2011)
- Campos JL, et al. *Molecular evolution in nonrecombining regions of the Drosophila melanogaster genome*. Genome Biol Evol. 4:278–288 (2012)
- Campos JL, et al. *The relation between recombination rate and patterns of molecular evolution and variation in Drosophila melanogaster*. Mol Biol Evol. 31: 1010–1028 (2014).
- Caracristi G and Schlötterer C. *Genetic differentiation between American and European Drosophila melanogaster populations could be attributed to admixture of African alleles*. Mol Biol Evol. 20(5): 792–799 (2003)
- Cargill M, et al. *Characterization of single-nucleotide polymorphisms in coding regions of human genes*. Nature Genet. 22(3): 231–238 (1999)
- Carlini DB and Stephan W. *In vivo introduction of unpreferred synonymous codons into the Drosophila Adh gene results in reduced levels of ADH protein*. Genetics 163: 239–243 (2003)
- Carroll SB. *Endless forms: the evolution of gene regulation and morphological diversity*. Cell 101: 577–580 (2000)
- Carroll SB. *Evolution at two levels: on genes and form*. PLoS Biol. 3: 1159 (2005)

- Carvalho AB, et al. *Y chromosome and other heterochromatic sequences of the Drosophila melanogaster genome: how far can we go?* Genetica 117(2-3): 227–237 (2003)
- Casillas S and Barbadilla A. *PDA v.2: improving the exploration and estimation of nucleotide polymorphism in large datasets of heterogeneous DNA.* Nucleic Acids Res. 34: W632-634 (2006)
- Casillas S, et al. *Purifying selection maintains highly conserved noncoding sequences in Drosophila.* Mol Biol Evol. 24(10): 2222–2234 (2007)
- Castellano D, et al. *Adaptive evolution is substantially impeded by Hill-Robertson interference in Drosophila.* Mol Biol Evol. 33(2): 442–455 (2016)
- Cavalli-Sforza LL. *Population structure and human evolution.* Proc R Soc Lond. B 164: 362–379 (1966)
- Celniker SE et al. *Finishing a whole-genome shotgun: release 3 of the Drosophila melanogaster euchromatic genome sequence.* Genome Biol. 3: RESEARCH0079 (2002)
- Chao L and Carr DE. *The molecular clock and the relationship between population size and generation time.* Evolution 47: 688–690 (1993)
- Charlesworth B, et al. *The relative rates of evolution of sex chromosomes and autosomes.* Am Nat. 130: 113–146 (1987)
- Charlesworth B, et al. *The distribution of transposable elements within and between chromosomes in a population of Drosophila melanogaster. II. Inferences on the nature of selection against elements.* Genet Res. 60: 115–30 (1992)
- Charlesworth B, et al. *The effect of deleterious mutations on neutral molecular variation.* Genetics 134: 1289–1303 (1993)
- Charlesworth B. *The effect of background selection against deleterious mutations on weakly selected, linked variants.* Genet Res. 63: 213–227 (1994)
- Charlesworth B, et al. *The effects of local selection, balanced polymorphism and background selection on equilibrium patterns of genetic diversity in subdivided populations.* Genet Res. 70: 155–174 (1997)
- Charlesworth B. *Measures of divergence between populations and the effect of forces that reduce variability.* Mol Biol Evol. 15: 538–543 (1998)
- Charlesworth B. *The effect of life-history and mode of inheritance on neutral genetic variability.* Genet Res. 77: 153–166 (2001)
- Charlesworth B, et al. *Estimating the incidence of ancestral polymorphisms.* Genet Res. 86: 149–157 (2005)
- Charlesworth B. *Fundamental concepts in genetics: effective population size and patterns of molecular evolution and variation.* Nat Rev Genet. 10: 195–205 (2009)
- Charlesworth B and Charlesworth D. *Elements of evolutionary genetics.* Greenwood Village, CO: Roberts and Co. Publishers (2010)
- Charlesworth B. *The effects of deleterious mutations on evolution at linked sites.* Genetics 190(1): 5–22 (2012a)
- Charlesworth B. *The role of background selection in shaping patterns of molecular evolution and variation: evidence from variability on the Drosophila X chromosome.* Genetics 191: 233–46 (2012b)

- Charlesworth B. *Background selection 20 years on: The Wilhelmine E. Key 2012 invitational lecture*. J Hered. 104: 161–171 (2013a)
- Charlesworth B. *Stabilizing selection, purifying selection, and mutational bias in finite populations*. Genetics 194: 955–71 (2013b).
- Charlesworth D, et al. *The pattern of neutral molecular variation under the background selection model*. Genetics 141: 1619–1632 (1995)
- Charlesworth D and Wright SI. *Breeding systems and genome evolution*. Curr Opin Genet Dev. 11: 685–690 (2001)
- Charlesworth D. *Balancing selection and its effects on sequences in nearby genome regions*. PLoS Genet. 2: e64 (2006)
- Charlesworth J and Eyre-Walker A. *The rate of adaptive evolution in enteric bacteria*. Mol Biol Evol. 23: 1348–1356 (2006)
- Charlesworth J and Eyre-Walker A. *The McDonald-Kreitman test and slightly deleterious mutations*. Mol Biol Evol. 25(6): 1007–1015 (2008)
- Chimpanzee Sequencing and Analysis Consortium. *Initial sequence of the chimpanzee genome and comparison with the human genome*. Nature 437: 69–87 (2005)
- Chun S and Fay JC. *Evidence for hitchhiking of deleterious mutations within the human genome*. PLoS Genet. 7: e1002240 (2011)
- Clark AG, et al. *Inferring nonneutral evolution from human-chimp-mouse orthologous gene trios*. Science 302: 1960–1963 (2003)
- Cliften P, et al. *Finding functional features in Saccharomyces genomes by phylogenetic footprinting*. Science 301: 71–76 (2003)
- Clop A, et al. *A mutation creating a potential illegitimate microRNA target site in the myostatin gene affects muscularity in sheep*. Nat Genet. 38(7): 813–818 (2006)
- Comeron JM, et al. *Natural selection on synonymous sites is correlated with gene length and recombination in Drosophila*. Genetics 151: 239–249 (1999)
- Comeron JM and Kreitman M. *The correlation between intron length and recombination in Drosophila: dynamic equilibrium between mutational and selective forces*. Genetics 156: 1175–1190 (2000)
- Comeron JM and Kreitman M. *Population, evolutionary and genomic consequences of interference selection*. Genetics 161: 389–410 (2002)
- Comeron JM, et al. *The Hill-Robertson effect: evolutionary consequences of weak selection and linkage in finite populations*. Heredity 100: 19–31 (2008)
- Comeron JM, et al. *The many landscapes of recombination in Drosophila melanogaster*. PLoS Genet. 8: e1002905 (2012)
- Comeron JM. *Background selection as baseline for nucleotide variation across the Drosophila genome*. PLoS Genet. 10: e1004434 (2014)
- Connallon T. *Adaptive protein evolution of X-linked and autosomal genes in Drosophila: implications for faster-X hypotheses*. Mol Biol Evol. 24 (11): 2566–2572 (2007)
- Coop G. *Does linked selection explain the narrow range of genetic diversity across species?* bioRxiv  
<http://dx.doi.org/10.1101/042598> (2016)

- Corbett-Detig RB and Hartl DL. *Population genomics of inversion polymorphisms in Drosophila melanogaster*. PLoS Genet. 8: e1003056 (2012)
- Corbett-Detig RB, et al. *Natural selection constrains neutral diversity across a wide range of species*. PLOS Biol. 13: e1002112 (2015)
- Correns G. *Mendel's Regel über das Verhalten der Nachkommenschaft der Rassenbastarde*. Berichte der Deutsche Botanischen Gesellschaft 18: 158-67 (1900)
- Craig DW, et al. *Identification of genetic variants using barcoded multiplexed sequencing*. Nature methods 5(10): 887-893 (2008)
- Crisci JL, et al. *Recent progress in polymorphism-based population genetic inference*. J Hered. 103: 287-296 (2012)
- Crow JF and Kimura M. *An Introduction to Population Genetic Theory*. New York: Harper and Row (1970)
- Crow JF and Kimura M. *Evolution in sexual and asexual populations*. Am Nat. 99: 439-450 (1965)
- Crow JF. *Breeding structure of populations. II. Effective population number. Statistics and Mathematic in Biology*. (ed.) O Kempthorne, Iowa State College Press, Ames (1954)
- Crow JF and Simmons MJ. *The mutation load in Drosophila*. In The Genetics and Biology of Drosophila, (ed.) M. Ashburner, H. L. Carson, and J. L. Thompson. Academic Press, London. pp. 1-26 (1983)
- Cutter AD and Payseur BA. *Selection at linked sites in the partial selfer Caenorhabditis elegans*. Mol Biol Evol 20: 665-673 (2003)
- Cutter AD and Payseur BA. *Genomic signatures of selection at linked sites: unifying the disparity among species*. Nat Rev Genet. 14: 262-274 (2013)
- Cutter AD and Choi JY. *Natural selection shapes nucleotide polymorphism across the genome of the nematode Caenorhabditis briggsae*. Genome Res. 20: 1103-1111 (2010)
- Cutter AD, et al. *Molecular hyperdiversity and evolution in very large populations*. Mol Ecol. 22: 2074-2095 (2013)
- Darwin C and Wallace A. *On the tendency of species to form varieties; and on the perpetuation of varieties and species by natural means of selection*. J Proc Linn Soc London Zool. 3: 45-62 (1858)
- Darwin C. *On the Origin of Species*. London: John Murray (1859)
- David J and Capy P. *Genetic variation of Drosophila melanogaster natural populations*. Trends Genet. 4(4): 106-111 (1988)
- Davidson EH. *Genomic regulatory systems: development and evolution*. Academic Press; San Diego (2001)
- Dayyodov EV, et al. *Identifying a high fraction of the human genome to be under selective constraint using GERP++*. PLOS Comput. Biol. 6: e1001025 (2010)
- Denniston C. *Small population size and genetic diversity: implications for endangered species*. In: Endangered birds: management techniques for preserving threatened species (Temple SA, ed). Madison, Wisconsin: University of Wisconsin Press; pp. 281-289 (1978)

- Dermitzakis ET, et al. *Conserved non-genic sequences - an unexpected feature of mammalian genomes*. Nat Rev Genet. 6: 151–157 (2005)
- Dickerson RE. *The structure of cytochrome c and the rates of molecular evolution*. J Mol Evol. 1(1): 26–45 (1971)
- Dobzhansky T. *Genetics and the Origin of Species*. Columbia University Press, New York (1937)
- Dobzhansky T. *Genetics of the Evolutionary Process*. Columbia University Press (1970)
- Drosophila 12 Genomes Consortium. *Evolution of genes and genomes on the Drosophila phylogeny*. Nature 450: 203–218 (2007)
- Duchen P, et al. *Demographic inference reveals African and European admixture in the north American Drosophila melanogaster population*. Genetics 193(1): 291–301 (2013)
- Duret L and Mouchiroud D. *Expression pattern and, surprisingly, gene length shape codon usage in Caenorhabditis, Drosophila and Arabidopsis*. Proc Natl Acad Sci. USA 96: 4482–4487 (1999)
- Durinck S, et al. BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. Bioinformatics 21: 3439–3440 (2005)
- Dvorak J, et al. *Restriction fragment length polymorphism and divergence in the genomic regions of high and low recombination in self-fertilizing and cross-fertilizing Aegilops species*. Genetics 148: 423–434 (1998)
- Egea R, et al. *Standard and generalized McDonald-Kreitman test: a website to detect selection by comparing different classes of DNA sites*. Nucleic Acids Res. 36: W157–162 (2008)
- Elyashiv E, et al. *A genomic map of the effects of linked selection in Drosophila*. arXiv <http://arxiv.org/abs/1408.5461> (2014)
- Emberly E, et al. *Conservation of regulatory elements between two species of Drosophila*. BMC Bioinformatics 4: 57–67 (2003)
- Enard D, et al. *Genome-wide signals of positive selection in human evolution*. Genome Res. 24: 885–895 (2014)
- Escalante AE, et al. *El estudio de la biodiversidad en la era de la secuenciación masiva*. Rev Mex Biodivers. 85: 1249–1264 (2014)
- Ewens WJ. *The sampling theory of selectively neutral alleles*. Theor Popul Biol. 3(1): 87–112 (1972)
- Ewens WJ. *Mathematical Population Genetics*. Berlin: Springer-Verlag (1979)
- Excoffier L, et al. *Detecting loci under selection in a hierarchically structured population*. Heredity 103(4): 285–298 (2009)
- Eyre-Walker A, et al. *Quantifying the slightly deleterious mutation model of molecular evolution*. Mol Biol Evol. 19: 2142–2149 (2002)
- Eyre-Walker A. *Changing effective population size and the McDonald-Kreitman test*. Genetics 162: 2017–2024 (2002)
- Eyre-Walker A, et al. *The distribution of fitness of new deleterious amino acid mutations in humans*. Genetics 173: 891–900 (2006)
- Eyre-Walker A and Keightley PD. *The distribution of fitness effects of new mutations*. Nat Rev Genet. 8: 610–618 (2007)

- Eyre-Walker A and Keightley PD. *Estimating the rate of adaptive molecular evolution in the presence of slightly deleterious mutations and population size change*. Mol Biol Evol 2: 2097–2108 (2009)
- Eyre-Walker A. *Evolution in health and medicine Sackler colloquium: Genetic architecture of a complex trait and its implications for fitness and genome-wide association studies*. Proc Natl Acad Sci. USA 107: 1752–1756 (2010)
- Falconer DS. *Introduction to Quantitative Genetics*. Edinburgh: Oliver and Boyd (1960)
- Falconer DS and Mackay TFC. *Introduction to Quantitative Genetics*. New York: Longman (1996)
- Fariello MI, et al. *Detecting signatures of selection through haplotype differentiation among hierarchically structured populations*. Genetics 193: 929–941 (2013)
- Fay JC and Wu CI. *Hitchhiking under positive Darwinian selection*. Genetics 155(3): 1405–1413 (2000)
- Fay JC, et al. *Positive and negative selection on the human genome*. Genetics 158: 1227–1234 (2001)
- Fay JC, et al. *Testing the neutral theory of molecular evolution with genomic data from Drosophila*. Nature 415: 1024–1026 (2002)
- Fay JC. *Weighing the evidence for adaptation at the molecular level*. Trends Genet. 27(9): 343–349 (2011)
- Feder AF, et al. *LDx: estimation of linkage disequilibrium from high-throughput pooled resequencing data*. PLoS One 7: e48588 (2012)
- Felsenstein J. *The evolutionary advantage of recombination*. Genetics 78: 737–756 (1974)
- Fisher RA. *The distribution of gene ratios for rare mutations*. Proc R Soc Edinburgh 50: 205–220 (1930)
- Fiston-Lavier AS, et al. *Drosophila melanogaster recombination rate calculator*. Gene 463: 18–20 (2010)
- Ford EB. *Ecological genetics*. 3. London, UK: Chapman and Hall (1971)
- Foss E, et al. *Chiasma interference as a function of genetic-distance*. Genetics 133(3): 681–691 (1993)
- Fox JA, et al. *A compilation of molecular biology web servers: 2006 update on the Bioinformatics Links Directory*. Nucleic Acids Res. 34: W3–5 (2006)
- Frankham R. *Effective population size/adult population size ratios in wildlife: a review*. Genet Res. 66: 95–107 (1995)
- Fu YX. *Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection*. Genetics 147(2): 915–925 (1997)
- Fu YX and Li WH. *Statistical tests of neutrality of mutations*. Genetics 133(3): 693–709 (1993)
- Fu W, et al. *Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants*. Nature 493: 216–220 (2013)
- Gaffney DJ and Keightley PD. *Genomic selective constraints in murid noncoding DNA*. PLoS Genet 2: e204 (2006)
- Gallo SM, et al. *REDfly v3.0: toward a comprehensive database of transcriptional regulatory elements in Drosophila*. Nucleic Acids Res. 39: D118–123 (2011)

- Galperin MY. *The Molecular Biology Database Collection: 2007 update*. Nucleic Acids Res. 35: D3–4 (2007)
- Galtier N. *Adaptive Protein Evolution in Animals and the Effective Population Size Hypothesis*. PLoS Genet. 12: e1005774 (2016)
- Gan X, et al. *Multiple reference genomes and transcriptomes for Arabidopsis thaliana*. Nature 477: 419–423 (2011)
- García-Dorado A and Caballero A. *On the average coefficient of dominance of deleterious spontaneous mutations*. Genetics 155: 1991–2001 (2000)
- García-Dorado A, et al. *Rates and effects of deleterious mutations and their evolutionary consequences*. In Moya A, Font E, (ed.) Evolution of molecules and ecosystems. Oxford: Oxford University Press. pp. 20–32 (2004)
- García-Dorado A. *Understanding and predicting the fitness decline of shrunk populations: inbreeding, purging, mutation, and standard selection*. Genetics 190(4): 1461–1476 (2012)
- Garud NR, et al. *Recent selective sweeps in North American Drosophila melanogaster show signatures of soft sweeps*. PLoS Genet. 11: e1005004 (2015)
- Gillespie JH. *The Causes of Molecular Evolution*. Oxford University Press, Oxford (1991)
- Gillespie JH. *Genetic Drift in an Infinite Population: The Pseudohitchhiking Model*. Genetics 155(2): 909–919 (2000a)
- Gillespie JH. *The neutral theory in an infinite population*. Genetics 261(1): 11–18 (2000b)
- Gillespie JH. *Is the population size of a species relevant to its evolution?* Evolution 55: 2161–2169 (2001)
- Goffeau A, et al. *Life with 6000 genes*. Science 274(5287): 546: 563–7 (1995)
- Golding GB. *The effect of purifying selection on genealogies*. In Progress in Population Genetics and Human Evolution, (ed.) P. Donnelly and S. Tavaré. Springer-Verlag, New York, pp. 271–285 (1997)
- Goldman N and Yang Z. *A codon based model of nucleotide substitution for protein-coding DNA sequences*. Mol Biol Evol. 11: 725–736 (1994)
- Gossmann TI, et al. *Genome wide analyses reveal little evidence for adaptive evolution in many plant species*. Mol Biol Evol. 27(8): 1822–1832 (2010)
- Gossmann TI, et al. *The effect of variation in the effective population size on the rate of adaptive molecular evolution in eukaryotes*. Genome Biol Evol. 4(5):658–67 (2012)
- Graur D and Li W-H. *Fundamentals of Molecular Evolution*. Sunderland, MA: Sinauer Assoc. (2000)
- Haddrill PR, et al. *Multilocus patterns of nucleotide variability and the demographic and selection history of Drosophila melanogaster populations*. Genome Res. 15 (6): 790–799 (2005)
- Haddrill PR, et al. *Reduced efficacy of selection in regions of the Drosophila genome that lack crossing over*. Genome Biol. 8: R18 (2007)
- Haddrill PR, et al. *Positive and negative selection on noncoding DNA in Drosophila simulans*. Mol Biol Evol. 25: 1825–1834 (2008)

- Haddrill PR, et al. *Estimating the parameters of selection on nonsynonymous mutations in Drosophila pseudoobscura and D. miranda*. Genetics 185: 1381–1396 (2010)
- Haddrill PR, et al. *Determinants of synonymous and nonsynonymous variability in three species of Drosophila*. Mol Biol Evol. 28: 1731–1743 (2011)
- Hahn MW, et al. *The effects of selection against spurious transcription factor binding sites*. Mol Biol Evol. 20(6): 901–906 (2003)
- Hahn MW. *Toward a selection theory of molecular evolution*. Evolution 62: 255–265 (2008).
- Haigh J and Maynard Smith J. *Population size and protein variation in man*. Genet Res. 19: 73–89 (1972)
- Haldane JBS. *The cost of natural selection*. J Genet. 55: 511–524 (1957)
- Haldane JBS. *Disease and evolution*. In *Malaria: Genetic and Evolutionary Aspects* (ed.) KR Dronamraju, P Arese. New York: Springer, pp. 175–187 (2006)
- Halligan DL and Keightley PD. *Ubiquitous selective constraints in the Drosophila genome revealed by a genome-wide interspecies comparison*. Genome Res. 16: 875–884 (2006)
- Halligan DL, et al. *Positive and negative selection in murine ultra-conserved noncoding elements*. Mol Biol Evol. 28: 2651–2660 (2011)
- Halligan DL, et al. *Contributions of protein-coding and regulatory change to adaptive molecular evolution in murid rodents*. PLoS Genet. 9: e1003995 (2013)
- Han L and Abney M. *Using identity by descent estimation with dense genotype data to detect positive selection*. Eur J Hum Genet. 21(2): 205–211 (2012)
- Hanchard NA, et al. *Screening for recently selected alleles by analysis of human haplotype similarity*. Am J Hum Genet. 78(1): 153–59 (2006)
- Hardy GH. *Mendelian Proportions in a Mixed Population*. Science 28(706): 49–50 (1908)
- Harr B, et al. *Hitchhiking mapping: a population-based fine-mapping strategy for adaptive mutations in Drosophila melanogaster*. Proc Natl Acad Sci. USA 99: 12949–12954 (2002)
- Harris H. *Enzyme polymorphisms in man*. Proc R Soc Lond B Biol Sci 164(995): 298–310 (1966)
- Haerty W, et al. *Evolution in the fast lane: rapidly evolving sex-related genes in Drosophila*. Genetics 177: 1321–1335 (2007)
- Haag-Liautard C, et al. *Direct estimation of per nucleotide and genomic deleterious mutation rates in Drosophila*. Nature 445(7123): 82–85 (2007)
- Hartfield M, et al. *The role of advantageous mutations in enhancing the evolution of a recombination modifier*. Genetics 184: 1153–64 (2010)
- Hellmann I, et al. *A neutral explanation for the correlation of diversity with recombination rates in humans*. Am J Hum Genet. 72: 1527–1535 (2003)
- Hermission J and Pennings PS. *Soft sweeps: molecular population genetics of adaptation from standing genetic variation*. Genetics 169: 2335–2352 (2005)

- Hernandez RD, et al. *Context dependence, ancestral misidentification, and spurious signatures of natural selection*. Mol Biol Evol. 24: 1792–1800 (2007)
- Hernandez RD. *A flexible forward simulator for populations subject to selection and demography*. Bioinformatics 24(23): 2786–2787 (2008)
- Hernandez RD, et al. *Classic selective sweeps were rare in recent human evolution*. Science 331: 920–924 (2011)
- Hershberg R and Petrov DA. *Selection on codon bias*. Annu Rev Genet. 42: 287–299 (2008)
- Hey J and Kliman RM. *Interactions between natural selection, recombination and gene density in the genes of Drosophila*. Genetics 160: 595–608 (2002)
- Hill WG and Robertson A. *Linkage disequilibrium in finite populations*. Theor Appl Genet. 38: 226–231 (1968)
- Hilliker AJ and Chovnick A. *Further observations on intragenic recombination in Drosophila melanogaster*. Genet Res. 38(2): 281–96 (1981)
- Hilliker AJ, et al. *The effect of DNA sequence polymorphisms on intragenic recombination in the rosy locus of Drosophila melanogaster*. Genetics 129(3): 779–781 (1991)
- Hinds DA, et al. *Whole-genome patterns of common DNA variation in three human populations*. Science 307: 1072–1079 (2005)
- Hoekstra HE and Coyne JA. *The locus of evolution: evo devo and the genetics of adaptation*. Evolution 61: 995–1016 (2007)
- Hoskins RA, et al. *Heterochromatic sequences in a Drosophila whole-genome shotgun assembly*. Genome Biol. 3(12): RESEARCH0085 (2002)
- Hoskins RA, et al. *Genome-wide analysis of promoter architecture in Drosophila melanogaster*. Genome Res. 21(2): 182–192 (2011)
- Hoskins RA, et al. *The release 6 reference sequence of the Drosophila melanogaster genome*. Genome Res. 25(3): 445–458 (2015)
- Howard LO. *A contribution to the study of the insect fauna of human excrement*. Proc Washington Acad Sci. 2: 541–604 (1900)
- Hu TT, et al. *A second-generation assembly of the Drosophila simulans genome provides new insights into patterns of lineage-specific divergence*. Genome Res. 23: 89–98 (2013)
- Huang W, et al. *Natural variation in genome architecture among 205 Drosophila melanogaster Genetic Reference Panel lines*. Genome Res. 24: 1193–1208 (2014)
- Hubby JL and Lewontin RC. *A molecular approach to the study of genic heterozygosity in natural populations. I. The number of alleles at different loci in Drosophila pseudoobscura*. Genetics 54: 577–594 (1966)
- Hudson RR. *Properties of a neutral allele model with intragenic recombination*. Theor Popul Biol. 23: 183–201 (1983)
- Hudson RR, et al. *A test of neutral molecular evolution based on nucleotide data*. Genetics 116(1): 153–159 (1987)
- Hudson RR and Kaplan NL. *Deleterious background selection with recombination*. Genetics 141: 1605–1617 (1995)

- Hughes AL. *Evidence for abundant slightly deleterious polymorphisms in bacterial populations*. Genetics 169: 533–538 (2005)
- Hurst LD. *The Ka/Ks ratio: diagnosing the form of sequence evolution*. Trends Genet. 18(9): 486 (2002)
- Hutter S, et al. *Genome-wide DNA polymorphism analyses using VariScan*. BMC Bioinformatics 7: 409 (2006)
- Hutter S, et al. *Distinctly different sex ratios in African and European populations of Drosophila melanogaster inferred from chromosome-wide nucleotide polymorphism data*. Genetics 177: 469–480 (2007)
- Iles MM, et al. *Recombination can evolve in large finite populations given selection on sufficient loci*. Genetics 165: 2249–2258 (2003)
- Ingvarsson PK. *Natural selection on synonymous and nonsynonymous mutations shapes patterns of polymorphism in Populus tremula*. Mol Biol Evol. 27: 650–660 (2010)
- International HapMap Consortium. *A haplotype map of the human genome*. Nature 437: 1299–1320 (2005)
- International HapMap Consortium. *A second generation human haplotype map of over 3.1 million SNPs*. Nature 447: 661–678 (2007)
- Ivanov AI, et al. *Genes required for Drosophila nervous system development identified by RNA interference*. Proc Natl Acad Sci. USA 101 (46): 16216–16221 (2004)
- Jakobsson M, et al. *Genotype, haplotype and copy-number variation in worldwide human populations*. Nature 451: 998–1003 (2008)
- Jenkins DL, et al. *A test for adaptive change in DNA sequences controlling transcription*. Proc Biol Sci. 261: 203–207 (1995)
- Jensen JD, et al. *An approximate Bayesian estimator suggests strong, recurrent selective sweeps in Drosophila*. PLoS Genet. 4(9): e1000198 (2008)
- Jensen JD. *On reconciling single and recurrent hitchhiking models*. Genome Biol Evol. 1: 320–324 (2009)
- Johnson CW. *The distribution of some species of Drosophila*. J Ent. 20(6): 202–205 (1913)
- Johnson FM et al. *An analysis of polymorphisms among isozyme loci in dark and light Drosophila ananassae strains from American and Western Samoa*. Proc Natl Acad Sci. USA 56(1): 119–125 (1966)
- Johnson KP and Seger J. *Elevated rates of nonsynonymous substitution in island birds*. Mol Biol Evol. 18: 874–881 (2001)
- Johnson T and Barton NH. *The effect of deleterious alleles on adaptation in asexual populations*. Genetics 162(1): 395–411 (2002)
- Jones FC, et al. *The genomic basis of adaptive evolution in threespine sticklebacks*. Nature 484(7392): 55–61 (2012)
- Jones GH. *Chiasmata*. In Mitosis, (ed.) P. B. Moensa. Academic Press, New York. pp. 213–244 (1987)
- Jones N. *Computer science: The learning machines*. Nature 505(7482): 146–148 (2014)
- Jukes TH and Cantor CR. *Mammalian Protein Metabolism*. In Evolution of Protein Molecules, (ed.) HN Munro. Academic Press, New York, pp. 21–132 (1969)
- Kacser H and Burns JA. *The molecular basis of dominance*. Genetics 97: 639–666 (1981)

- Kaminker J, et al. *The transposable elements of the Drosophila melanogaster euchromatin: a genomics perspective*. Genome Biol. 3(12): RESEARCH0084 (2002)
- Karasov TT, et al. *Evidence that adaptation in Drosophila is not limited by mutation at single sites*. PLoS Genet. 6: e1000924 (2010)
- Kayserili MA, et al. *An excess of gene expression divergence on the X chromosome in Drosophila embryos: implications for the faster-X hypothesis*. PLoS Genet. 8: e1003200 (2012)
- Keightley PD, et al. *Evolutionary constraints in conserved nongenic sequences of mammals*. Genome Res. 15: 1373–1378 (2005a)
- Keightley PD, et al. *Evidence for widespread degradation of gene control regions in hominid genomes*. PLoS Biol. 3: e42. (2005b)
- Keightley PD and Eyre-Walker A. *Joint inference of the distribution of fitness effects of deleterious mutations and population demography based on nucleotide polymorphism frequencies*. Genetics 177: 2251–2261 (2007)
- Keightley PD and Otto SP. *Interference among deleterious mutations favours sex and recombination in finite populations*. Nature 443: 89–92 (2006)
- Keightley PD et al. *Analysis of the genome sequences of three Drosophila melanogaster spontaneous mutation accumulation lines*. Genome Res. 19: 1195–1201 (2009)
- Keightley PD and Eyre-Walker A. *What can we learn about the distribution of fitness effects of new mutations from DNA sequence data?* Philos Trans R Soc Lond B Biol Sci. 365: 1187–1193 (2010)
- Keightley PD and Eyre-Walker A. *Estimating the rate of adaptive molecular evolution when the evolutionary divergence between species is small*. J Mol Evol 74: 61–68 (2012)
- Keinan A and Reich D. *Human population differentiation is strongly correlated with local recombination rate*. PLoS Genet. 6: e1000886 (2010)
- Keller A. *Drosophila melanogaster's history as a human commensal*. Curr Biol. 17(3): R77–R81 (2007)
- Kim S, et al. *Recombination and linkage disequilibrium in Arabidopsis thaliana*. Nat Genet. 39: 1151–1155 (2007)
- Kimura M. *Some problems of stochastic processes in genetics*. Ann Math Statist. 28(4): 882–901 (1957)
- Kimura, M. *On the probability of fixation of mutant genes in a population*. Genetics 47(6): 713–719 (1962)
- Kimura M and Crow JF. *The number of alleles that can be maintained in a finite population*. Genetics 49: 725–738 (1964)
- Kimura M. *Evolutionary rate at the molecular level*. Nature 217: 624–626 (1968)
- Kimura M. *The rate of molecular evolution considered from the standpoint of population genetics*. Proc Natl Acad Sci. USA 63: 1181–1188 (1969a)
- Kimura M. *The number of heterozygous nucleotide sites maintained in a finite population due to the steady flux of mutations*. Genetics 61: 893–903 (1969b)
- Kimura M and Ohta T. *the average number of generations until fixation of a mutant gene in a finite population*. Genetics 61(3): 763–771 (1969)

- Kimura M. *Theoretical foundation of population genetics at the molecular level*. *Theor Pop Biol.* 2: 174–208 (1971)
- Kimura M and Ohta T. *On some principles governing molecular evolution*. *Proc Natl Acad Sci. USA* 71: 2848–2852 (1974)
- Kimura M. *Model of effectively neutral mutations in which selective constraint is incorporated*. *Proc Natl Acad Sci. USA* 76: 3440–3444 (1979)
- Kimura M. *The Neutral Theory of Molecular Evolution*. Cambridge University Press, Cambridge (1983)
- King JL and Jukes TH. *Non-Darwinian evolution*. *Science* 164: 788–798 (1969)
- King M-C and Wilson AC. *Evolution at two levels in humans and chimpanzees*. *Science* 188: 107–116 (1975)
- Kingman JFC. *On the genealogy of large populations*. *J Appl Probab.* 19: 27–43 (1982a)
- Kingman JFC. *The coalescent*. *Stochastic Process Appl.* 13(3): 235–248 (1982b)
- Kliman RM and Hey J. *Reduced natural selection associated with low recombination in Drosophila melanogaster*. *Mol Biol Evol.* 10: 1239–1258 (1993).
- Kofler R and Schlötterer C. *Low levels of transposable element activity in Drosophila mauritiana : causes and consequences*. bioRxiv <http://dx.doi.org/10.1101/018218> (2015)
- Kohn MH, et al. *Inference of positive and negative selection on the 5' regulatory regions of Drosophila genes*. *Mol Biol Evol* 21: 374–383 (2004)
- Kopczynski CC, et al. *A high throughput screen to identify secreted and transmembrane proteins involved in Drosophila embryogenesis*. *Proc Natl Acad Sci. USA* 95(17): 9973–9978 (1998)
- Kousathanas A, et al. *Positive and negative selection on non-coding DNA close to protein-coding genes in wild house mice*. *Mol Biol Evol.* 28(3): 1183–1191 (2011)
- Kousathanas A and Keightley PD. *A comparison of models to infer the distribution of fitness effects of new mutations*. *Genetics* 193: 1197–1208 (2013)
- Kraft T, et al. *Positive correlation between recombination rates and levels of genetic variation in natural populations of sea beet (*Beta vulgaris* subsp. *maritima*)*. *Genetics* 150: 1239–1244 (1998)
- Kreitman M. *Nucleotide polymorphism at the alcohol dehydrogenase locus of Drosophila melanogaster*. *Nature* 304 (5925): 412–417 (1983)
- Krone SM and Neuhauser C. *Ancestral processes with selection*. *Theor Popul Biol.* 51: 210–237 (1997)
- Kryukov GV, et al. *Small fitness effect of mutations in highly conserved non-coding regions*. *Hum Mol Genet.* 14, 2221–2229 (2005)
- Kuhner MK. *LAMARC 2.0: maximum likelihood and Bayesian estimation of population parameters*. *Bioinformatics* 22: 768–70 (2006)
- Kumar S. *Molecular clocks: four decades of evolution*. *Nat Rev Genet.* 6: 654–62 (2005)

- Kumar S, et al. MEGA-CC: computing core of molecular evolutionary genetics analysis program for automated and iterative data analysis. *Bioinformatics* 28: 2685–2686 (2012).
- Lachaise D, et al. Historical biogeography of *Drosophila melanogaster* species subgroup. *Evol Biol.* 22: 159–225 (1988)
- Lack JB, et al. The *Drosophila* genome nexus: a population genomic resource of 623 *Drosophila melanogaster* genomes, including 197 from a single ancestral range population. *Genetics* 199: 1229–1241 (2015)
- Lander ES, et al. Initial sequencing and analysis of the human genome. *Nature* 409: 860–921 (2001)
- Lanfear R, et al. Population size and the rate of evolution. *Trends Ecol Evol.* 29: 33–41 (2014).
- Langley CH, et al. On the role of unequal exchange in the containment of transposable element copy number. *Genet Res.* 52: 223–235 (1989)
- Langley CH, et al. Genomic variation in natural populations of *Drosophila melanogaster*. *Genetics* 192: 533–598 (2012)
- Lawrie DS, et al. Strong purifying selection at synonymous sites in *D. melanogaster*. *PLoS Genet.* 9: e1003527 (2013)
- Lee YCG, et al. Differential strengths of positive selection revealed by hitchhiking effects at small physical scales in *Drosophila melanogaster*. *Mol Biol Evol.* 31: 804–16 (2014)
- Leffler EM, et al. Revisiting an old riddle: what determines genetic diversity levels within species? *PLoS Biol.* 10: e1001388 (2012)
- Lesecque Y, et al. A resolution of the mutation load paradox in humans. *Genetics* 191(4): 1321–1330 (2012)
- Lewin B. *Genes IX*. Oxford University Press. pp. 892 (2007)
- Lewis EB. A Gene complex controlling segmentation in *Drosophila*. *Nature* 276(5688): 565–570 (1978)
- Lewontin RC and Hubby JL. A molecular approach to the study of genic heterozygosity in natural populations. II. Amount of variation and degree of heterozygosity in natural populations of *Drosophila pseudoobscura*. *Genetics* 54(2): 595–609 (1966)
- Lewontin RC and Krakauer J. Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. *Genetics* 74: 175–195 (1973)
- Lewontin RC. *The genetic basis of evolutionary change*. Columbia University Press, New York. (1974)
- Lewontin RC. *Biology as ideology: the doctrine of DNA*. Anansi Press / Stoddard Publishing Company (1992)
- Li H and Stephan W. Inferring the demographic history and rate of adaptive substitution in *Drosophila*. *PLoS Genet.* 2: e166 (2006)
- Li J, et al. Joint analysis of demography and selection in population genetics: where do we stand and where could we go? *Mol Ecol.* 21: 28–44 (2012).
- Li JZ, et al. Worldwide human relationships inferred from genome-wide patterns of variation. *Science* 319: 1100–1104 (2008)
- Li WH, et al. An evaluation of the molecular clock hypothesis using mammalian DNA sequences. *J Mol Evol.* 25: 330–342 (1987)

- Li WH. *Models of nearly neutral mutations with particular implications for nonrandom usage of synonymous codons*. J Mol Evol. 24: 337–345 (1987)
- Li WH. *Unbiased estimation of the rates of synonymous and non-synonymous substitution*. J Mol Evol. 36: 96–99 (1993)
- Li Y, et al. *Resequencing of 200 human exomes identifies an excess of low-frequency non-synonymous coding variants*. Nat Genet. 42(11): 969–972 (2010)
- Lin K, et al. *Distinguishing positive selection from neutral evolution: boosting the performance of summary statistics*. Genetics 187: 229–244 (2011)
- Lin K, et al. *A fast estimate for the population recombination rate based on regression*. Genetics 194: 473–484 (2013)
- Lindblad-Toh K, et al. *A high-resolution map of human evolutionary constraint using 29 mammals*. Nature 478: 476–482 (2011)
- Liti G, et al. *Population genomics of domestic and wild yeasts*. Nature 458(7236): 337–341 (2009)
- Llopart A. *The rapid evolution of X-linked male-biased gene expression and the large-X effect in Drosophila yakuba, D. santomea, and their hybrids*. Mol Biol Evol. 29: 3873–3886 (2012)
- Loewe L. *Quantifying the genomic decay paradox due to Muller's ratchet in human mitochondrial DNA*. Genet Res. 87: 133–159 (2006)
- Loewe L, et al. *Estimating selection on nonsynonymous mutations*. Genetics 172:1079–1092 (2006)
- Loewe L and Charlesworth B. *Inferring the distribution of mutational effects on fitness in Drosophila*. Biol Lett. 2: 426–430 (2006)
- Lohmueller KE, et al. *Natural selection affects multiple aspects of genetic variation at putatively neutral sites across the human genome*. PLoS Genet. 7: e1002326 (2011)
- Lourenço JM, et al. *The rate of molecular adaptation in a changing environment*. Mol Biol Evol. 30(6): 1292–1301 (2013)
- Luikart G, et al. *Distortion of allele frequency distributions provides a test for recent population bottlenecks*. J Hered. 89: 238–247 (1998)
- Lupski JR. *Genomic disorders ten years on*. Genome Med. 1(4): 42 (2009)
- Lynch M and Lande R. *The critical effective size for a genetically secure population*. Animal Conservation 1: 70–72 (1998)
- Lynch M. *The origins of genome architecture*. Sunderland, MA: Sinauer Associates (2007)
- Lynch M. *Evolution of the mutation rate*. Trends Genet. 26: 345–52 (2010)
- Lynch M. *The lower bound to the evolution of mutation rates*. Genome Biol Evol. 3: 1107–1118 (2011)
- MacArthur RH and Wilson EO. *The Theory of Island Biogeography*. Princeton Univ. Press (1967)
- Macpherson JM, et al. *Genomewide spatial correspondence between nonsynonymous divergence and neutral polymorphism reveals extensive adaptation in Drosophila*. Genetics 177: 2083–2099 (2007)
- Mackay TFC, et al. *The Drosophila melanogaster Genetic Reference Panel*. Nature 482: 173–178 (2012)

- Maruyama T and Fuerst PA. *Population bottlenecks and non-equilibrium models in population genetics. II. Number of alleles in a small population that was formed by a recent bottleneck*. Genetics 111: 675–689 (1985)
- Maside X and Charlesworth B. *Patterns of molecular variation and evolution in Drosophila americana and its relatives*. Genetics 176: 2293–2305 (2007)
- Matthews KA, et al. *Research resources for Drosophila: the expanding universe*. Nat Rev Genet. 6(3): 179–193 (2005)
- Maxam AM and Gilbert W. *A new method for sequencing DNA*. Proc Natl Acad Sci. USA 74(2): 560–564 (1977)
- Maynard Smith J and Haigh J. *The hitch-hiking effect of a favourable gene*. Genet Res. 23: 23–35 (1974)
- Maynard Smith J. *Evolution in sexual and asexual populations*. Am Nat. 102: 469–473 (1968)
- Mayr E. *Systematics and the Origin of Species*. Columbia University Press, New York (1942)
- McDonald JH and Kreitman M. *Adaptive protein evolution at the Adh locus in Drosophila*. Nature 351(6328): 652–654 (1991)
- McGaugh SE, et al. *Recombination modulates how selection affects linked sites in Drosophila*. PLoS Biol. 10: e1001422 (2012)
- McVicker G, et al. *Widespread genomic signatures of natural selection in hominid evolution*. PLoS Genet. 5: e1000471 (2009)
- McVean GA and Charlesworth B. *The effects of Hill-Robertson interference between weakly selected mutations on patterns of molecular evolution and variation*. Genetics 155: 929–944 (2000)
- McVean GA and Vieira J. *Inferring parameters of mutation, selection and demography from patterns of synonymous site evolution in Drosophila*. Genetics 157(1931): 245–257 (2001)
- Meiklejohn CD, et al. *Rapid evolution of male-biased gene expression in Drosophila*. Proc Natl Acad Sci USA. 100: 9894–9899 (2003)
- Meisel RP, et al. *Faster-X evolution of gene expression in Drosophila*. PLoS Genet. 8: e1003013 (2012)
- Meisel RP and Connallon T. *The faster-X effect: integrating theory and data*. Trends Genet. 29: 537–544 (2013)
- Mendel G. *Versuche über Pflanzen-Hybriden*. Verh. Naturforsch. Ver. Brünn 4: 3–47 (1866)
- Messer PW and Petrov DA. *Frequent adaptation and the McDonald-Kreitman test*. Proc Natl Acad Sci. USA 110: 8615–8620 (2013)
- Misra S, et al. *Annotation of the Drosophila melanogaster euchromatic genome: a systematic review*. Genome Biol. 3 (12): RESEARCH0083 (2002)
- Montgomery EA, et al. *A test for the role of natural selection in the stabilization of transposable element copy number in a population of Drosophila melanogaster*. Genet Res. 49: 31–41 (1987)
- Mooers AO and Harvey PH. *Metabolic rate, generation time, and the rate of molecular evolution in birds*. Mol Phylogenetic Evol. 3: 344–350 (1994)
- Morgan TH, et al. *The Mechanism of Mendelian Heredity*. Henry Holt & Company, New York (1915)

- Moriyama EN and Hartl DL. *Codon usage bias and base composition of nuclear genes in Drosophila*. Genetics 134: 847–858 (1993)
- Mortimer RK and Fogel S. *Genetical Interference and Gene Conversion*. In Mechanism in Recombination, (ed.) RF Grell, New York. pp. 263–275 (1974)
- Mouse Genome Sequencing Consortium. *Initial sequencing and comparative analysis of the mouse genome*. Nature 420: 520–562 (2002)
- Mukai T. *The genetic structure of natural populations of Drosophila melanogaster. I. Spontaneous mutation rate of polygenes controlling viability*. Genetics 50: 1–19 (1964)
- Mullen LM, et al. *Adaptive basis of geographic variation: genetic, phenotypic and environmental differences among beach mouse populations*. Proc Biol Sci. 276(1674): 3809–3818 (2009)
- Muller HJ. *Artificial transmutation of the gene*. Science 66(1699): 84–87 (1927)
- Muller HJ and Kaplan WD. *The dosage compensation of Drosophila and mammals as showing the accuracy of the normal type*. Genetical Research 8 (01): 41–59 (1966)
- Mullis KB and Faloona FA. *Specific synthesis of DNA in vitro via a polymerase-catalyzed chain reaction*. Methods Enzymol. 155: 335–50 (1987)
- Nabholz B, et al. *Determination of mitochondrial genetic diversity in mammals*. Genetics 178: 351–361 (2008)
- Nachman MW. *Patterns of DNA variability at X-linked loci in Mus domesticus*. Genetics 147: 1303–1316 (1997)
- Nachman MW. *Deleterious mutations in animal mitochondrial DNA*. Genetica 102: 61–69 (1998)
- Nachman MW, et al. *DNA variability and recombination rates at X-linked loci in humans*. Genetics 150: 1133–1141 (1998)
- Navarro A, et al. *Dynamics of gametic disequilibria between loci linked to chromosome inversions: the recombination redistributing effect of inversions*. Genet Res. 67(1): 67–76 (1996)
- Nei M, et al. *Infinite allele model with varying mutation rate*. Proc Natl Acad Sci. USA 73: 4164–4168 (1976)
- Nei M, et al. *The neutral theory of molecular evolution in the genomic era*. Annu Rev Genomics Hum Genet. 11, 265–89 (2010)
- Nelson CE, et al. *The regulatory content of intergenic DNA shapes genome architecture*. Genome Biol. 5: R25 (2004)
- Nelson MR, et al. *An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people*. Science 337: 100–104 (2012)
- Neuhauser C and Krone SM. *The genealogy of sample in models with selection*. Genetics 145: 519–534 (1997)
- Nielsen R and Yang Z. *Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene*. Genetics 148: 929–936 (1998)
- Nielsen R and Weinreich DM. *The age of nonsynonymous and synonymous mutations in animal mtDNA and implications for the mildly deleterious theory*. Genetics 153: 497–506 (1999)

- Nielsen R, et al. *Genomic scans for selective sweeps using SNP data*. *Genome Res.* 15: 1566–1575 (2005a)
- Nielsen R, et al. *A scan for positively selected genes in the genomes of humans and chimpanzees*. *PLoS Biol.* 3: e170 (2005b)
- Nordborg M. *Linkage disequilibrium, gene trees and selfing: an ancestral recombination graph with partial self-fertilization*. *Genetics* 154: 923–929 (2000)
- Nordborg M, et al. *The pattern of polymorphism in Arabidopsis thaliana*. *PLoS Biol.* 3: e196 (2005)
- Norton HL, et al. *Genetic evidence for the convergent evolution of light skin in Europeans and East Asians*. *Mol Biol Evol.* 24(3): 710–722 (2007)
- Noor MAF and Bennett SM. *Islands of speciation or mirages in the desert? Examining the role of restricted recombination in maintaining species*. *Heredity* 103: 439–444 (2009)
- Nüsslein-Volhard C and Wieschaus E. *Mutations affecting segment number and polarity in Drosophila*. *Nature* 287(5785): 795–801 (1980)
- Obbard DJ, et al. *Quantifying adaptive evolution in the Drosophila immune system*. *PLoS Genet.* 5: e1000698 (2009)
- Ohta T and M. Kimura M. *On the constancy of the evolutionary rate of cistrons*. *J Mol Evol.* 1: 18–25 (1971)
- Ohta T. *Evolutionary rate of cistrons and DNA divergence*. *J Mol Evol.* 1: 150–157 (1972a)
- Ohta T. *Population size and rate of evolution*. *J Mol Evol.* 1: 305–314 (1972b)
- Ohta T. *Slightly deleterious mutant substitutions in evolution*. *Nature* 246: 96–98 (1973)
- Ohta T. *Role of slightly deleterious mutations in molecular evolution and polymorphism*. *Theor Popul Biol.* 10: 254–275 (1976)
- Ohta T. *Extension of the neutral mutation drift hypothesis*. In M. Kimura, (ed.) *Molecular evolution and polymorphism*. National Institute of Genetics, Mishima, Japan. pp. 148–167 (1977)
- Ohta T. *The nearly neutral theory of molecular evolution*. *Annu Rev Ecol Syst.* 23: 263–286 (1992)
- Ohta T. *Amino acid substitution at the Adh locus of Drosophila is facilitated by small population size*. *Proc Natl Acad Sci. USA* 90: 4548–4551 (1993)
- Ohta T. *Synonymous and nonsynonymous substitutions in mammalian genes and the nearly neutral theory*. *J Mol Evol* 40: 56–63 (1995)
- Orr HA and Betancourt AJ. *Haldane's sieve and adaptation from the standing genetic variation*. *Genetics* 157: 875–84 (2001)
- Otto SP and Day T. *A biologist's guide to mathematical modelling in ecology and evolution*. Princeton University Press, Princeton (2007)
- Parmley JL, et al. *Evidence for purifying selection against synonymous mutations in mammalian exonic splicing enhancers*. *Mol Biol Evol.* 23: 301–309 (2006)
- Parsch J, et al. *Site-directed mutations reveal long range compensatory interactions in the Adh gene of Drosophila melanogaster*. *Proc Natl Acad Sci. USA* 94: 928–933 (1997)

- Parsch J, et al. *On the utility of short intron sequences as a reference for the detection of positive and negative selection in Drosophila*. Mol Biol Evol. 27:1226–1234 (2010)
- Pavlidis P, et al. *Searching for footprints of positive selection in whole-genome SNP data from nonequilibrium populations*. Genetics 185: 907–922 (2010)
- Pavlidis P, et al. *SweeD: likelihood-based detection of selective sweeps in thousands of genomes*. Mol Biol Evol. 30: 2224–2234 (2013)
- Peck JR, et al. *Imperfect genes, Fisherian mutation and the evolution of sex*. Genetics 145: 1171–1199 (1997)
- Peden JF. *Analysis of codon usage* [PhD thesis]. [Nottingham (UK)]: University of Nottingham. CodonW: Correspondence analysis of codon usage. Available from: <http://codonw.sourceforge.net/> (1999)
- Peter BM, et al. *Distinguishing between selective sweeps from standing variation and from a de novo mutation*. PLoS Genet. 8(10): e1003011 (2012)
- Piganeau G and Eyre-Walker A. *Evidence for variation in the effective population size of animal mitochondrial DNA*. PLoS One 4(2): e4396 (2009)
- Pollard KS, et al. *Forces shaping the fastest evolving regions in the human genome*. PLoS Genet. 2(10): e168 (2006)
- Pool JE and Nielsen R. *Population size changes reshape genomic patterns of diversity*. Evolution 61: 3001–3006 (2007)
- Pool JE and Nielsen R. *The impact of founder events on chromosomal variability in multiply mating species*. Mol Biol Evol. 25: 1728–1736 (2008)
- Pool JE, et al. *Population genetic inference from genomic sequence variation*. Genome Res. 20: 291–300 (2010)
- Pool JE, et al. *Population genomics of sub-Saharan Drosophila melanogaster: African diversity and non-African admixture*. PLoS Genet. 8: e1003080 (2012)
- Powell JR. *Progress and Prospects in Evolutionary Biology: The Drosophila Model*. Oxford University Press, New York (1997)
- Prabhakar S, et al. *Accelerated evolution of conserved noncoding sequences in humans*. Science 314(5800): 786 (2006)
- Presgraves DC. *Recombination enhances protein adaptation in Drosophila melanogaster*. Curr Biol. 15: 1651–1656 (2005)
- Pritchard JK, et al. *The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation*. Curr Biol. 20: R208–15 (2010)
- Pritchard JK and Di Rienzo A. *Adaptation - not by sweeps alone*. Nat Rev Genet. 11:665–7 (2010)
- Pröschel M, et al. *Widespread adaptive evolution of Drosophila genes with sex-biased expression*. Genetics 174:893–900 (2006)
- Przeworski M, et al. *Genealogies and weak purifying selection*. Mol Biol Evol. 16: 246–252 (1999)
- Przeworski M, et al. *Adjusting the focus on human variation*. Trends Genet. 16: 296–302 (2000)
- Przeworski M, et al. *The signature of positive selection on standing genetic variation*. Evolution 59: 2312–2323 (2005)

- Pybus M, et al. *Hierarchical boosting: a machine-learning framework to detect and classify hard selective sweeps in human populations*. Bioinformatics 31: 3946–3952 (2015)
- R Core Team. *R: a language and environment for statistical computing*. Vienna (Austria): R Foundation for Statistical Computing. ISBN 3-900051-07-0, Available from: <http://www.R-project.org/> (2013)
- Ràmia M, et al. *PopDrowser: The population Drosophila browser*. Bioinformatics 28(4): 595–596 (2012)
- Ràmia M. *Visualization and analysis of the genome of a natural population of D. melanogaster* [PhD thesis]. [Barcelona [Spain]]: Universitat Autònoma de Barcelona. Available from: <http://hdl.handle.net/10803/322822> (2015)
- Ramirez-Soriano A, et al. *Statistical power analysis of neutrality tests under demographic expansions, contractions and bottlenecks with recombination*. Genetics 179: 555–567 (2008)
- Rand DM and Kann LM. *Excess amino acid polymorphism in mitochondrial DNA: contrasts among genes from Drosophila, mice, and humans*. Mol Biol Evol. 13(6): 735–748 (1996)
- Razeto-Barry P, et al. *The nearly neutral and selection theories of molecular evolution under the fisher geometrical framework: substitution rate, population size, and complexity*. Genetics 191(2): 523–34 (2012)
- Reed FA, et al. *Fitting back-ground-selection predictions to levels of nucleotide variation and divergence along the human autosomes*. Genome Res. 15: 1211–1221 (2005)
- Roberts DB. *Drosophila melanogaster: The Model Organism*. Entomologia Experimentalis et Applicata 121(2): 93–103 (2006)
- Robertson A. *Inbreeding in artificial selection programmes*. Genet Res. 2: 189–194 (1961)
- Romiguier J, et al. *Comparative population genomics in animals uncovers the determinants of genetic diversity*. Nature 515: 261–263 (2014)
- Ronen R, et al. *Learning natural selection from the site frequency spectrum*. Genetics 195: 181–193 (2013)
- Roselius K, et al. *The relationship of nucleotide polymorphism, recombination rate and selection in wild tomato species*. Genetics 171: 753–763 (2005)
- Roze D and Barton NH. *The Hill–Robertson effect and the evolution of recombination*. Genetics 173: 1793–1811 (2006)
- Rubin GM. *Around the genomes: The Drosophila genome project*. Genome Res. 6(2): 71–79 (1996)
- Russo CA, et al. *Molecular phylogeny and divergence times of Drosophilid species*. Mol Biol Evol. 12(3): 391–404 (1995)
- Sabeti PC, et al. *Detecting recent positive selection in the human genome from haplotype structure*. Nature 419(6909): 832–837 (2002)
- Sabeti PC, et al. *Positive natural selection in the human lineage*. Science 312(5780): 1614–1620 (2006)
- Sabeti PC, et al. *Genome-wide detection and characterization of positive selection in human populations*. Nature 449(7164): 913–918 (2007)
- Sackton TB, et al. *Population genomic Inferences from sparse high-throughput sequencing of two populations of Drosophila melanogaster*. Genome Biol Evol. 1: 449–465 (2009)

- Sanger F and Coulson AR. *A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase*. J Mol Biol. 94(3): 441–448 (1975)
- Sattath S, et al. *Pervasive adaptive protein evolution apparent in diversity patterns around amino acid substitutions in Drosophila simulans*. PLoS Genet. 7: e1001302 (2011)
- Sawyer SA, et al. *Confidence interval for the number of selectively neutral amino acid polymorphisms*. Proc Natl Acad Sci. USA 84: 6225–6228 (1987)
- Sawyer SA, et al. *Bayesian analysis suggests that most amino acid replacements in Drosophila are driven by positive selection*. J Mol Evol. 57: S154–S164 (2003)
- Schneider A, et al. *A method for inferring the rate of occurrence and fitness effects of advantageous mutations*. Genetics 189(4): 1427–1437 (2011)
- Schrider DR, et al. *Rates and genomic consequences of spontaneous mutational events in Drosophila melanogaster*. Genetics 194(4): 937–954 (2013)
- Schultz ST and Lynch M. *Mutation and extinction: the role of variable mutational effects, synergistic epistasis, beneficial mutations and degree of outcrossing*. Evolution 51: 1363–1371 (1997)
- Sella G, et al. *Pervasive natural selection in the Drosophila genome?* PLoS Genet. 5: e1000495 (2009)
- Shabalina SA and Kondrashov AS. *Pattern of selective constraint in C. elegans and C. briggsae genomes*. Genet Res. 74: 23–30 (1999)
- Shapiro BJ and Alm EJ. *Comparing patterns of natural selection across species using selective signatures*. PLoS Genet. 4(2): e23 (2008)
- Sharp PM and Li WH. *The codon adaptation index—a measure of directional synonymous codon usage bias, and its potential applications*. Nucleic Acids Res. 15: 1281–1295 (1987)
- Sharp PM and Li WH. *On the rate of DNA sequence evolution in Drosophila*. J Mol Evol. 28(5): 398–402 (1989)
- Shen H, et al. *Comprehensive characterization of human genome variation by high coverage whole-genome sequencing of forty four Caucasians*. PLoS One 8: e59494 (2013)
- Shields DC, et al. *“Silent” sites in Drosophila genes are not neutral: evidence of selection among synonymous codons*. Mol Biol Evol. 5:704–716 (1988)
- Shriver MD, et al. *The genomic distribution of population substructure in four populations using 8,525 autosomal SNPs*. Hum Genomics 1(4): 274–286 (2004)
- Siepel A, et al. *Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes*. Genome Res. 15: 1034–1050 (2005)
- Simonsen KL, et al. *Properties of statistical tests of neutrality for DNA polymorphism data*. Genetics 141: 413–429 (1995)
- Simpson GG. *Tempo and Mode in Evolution*. Columbia University Press, New York (1944)
- Singh ND, et al. *A. Similar levels of X-linked and autosomal nucleotide variation in African and non-African populations of Drosophila melanogaster*. BMC Evol Biol. 7: 202 (2007)

- Singh ND, et al. *Contrasting the efficacy of selection on the X and autosomes in Drosophila*. Mol Biol Evol. 25: 454–67 (2008)
- Singh RS and Rhomberg LR. *A comprehensive study of genic variation in natural populations of Drosophila melanogaster. I. Estimates of gene flow from rare alleles*. Genetics 115(2): 313–322 (1987)
- Sironi M, et al. *Analysis of intronic conserved elements indicates that functional complexity might represent a major source of negative selection on non-coding sequences*. Hum Mol Genet. 14: 2533–2546 (2005)
- Slotman MA, et al. *Reduced recombination rate and genetic differentiation between the M and S forms of Anopheles gambiae s.s.* Genetics 174: 2081–2093 (2006)
- Slotte T, et al. *Genome-wide evidence for efficient positive and purifying selection in Capsella grandiflora, a plant species with a large effective population size*. Mol Biol Evol. 27: 1813–1821 (2010)
- Smith DR, et al. *Rapid whole-genome mutational profiling using next-generation sequencing technologies*. Genome Res. 18(10): 1638–1642 (2008)
- Smith LM, et al. *Fluorescence detection in automated DNA sequence analysis*. Nature 321 (6071): 674–679 (1986)
- Smith NGC and Eyre-Walker A. *Adaptive protein evolution in Drosophila*. Nature 415: 1022–1024 (2002)
- Spiess EB. *Genes in Populations*. 2nd editio. John Wiley & Sons, New York (1989)
- Staden R. *A strategy of DNA sequencing employing computer programs*. Nucleic Acids Res. 6(7): 2601–2610 (1979)
- Stajich JE and Hahn MW. *Disentangling the effects of demography and selection in human history*. Mol Biol Evol. 22: 63–73 (2005)
- Stebbins GL. *Variation and Evolution in Plants*. Columbia University Press, New York (1950)
- Stephan W and Langley CH. *Evolutionary consequences of DNA mismatch inhibited repair opportunity*. Genetics 132: 567–74 (1992)
- Stephan W and Langley CH. *DNA polymorphism in Lycopersicon and crossing-over per physical length*. Genetics 150: 1585–1593 (1998)
- Stephan W, et al. *The effect of background selection at a single locus on weakly selected, partially linked variants*. Genet Res. 73:133–146 (1999)
- Stephan W and Li H. *The recent demographic and adaptive history of Drosophila melanogaster*. Heredity 98(2): 65–68 (2007)
- Stephan W. *Genetic hitchhiking versus background selection: the controversy and its implications*. Philos Trans R Soc Lond B Biol Sci. 365: 1245–1253 (2010)
- Stern DL. *Evolution, development and the predictable genome*. Roberts and Co. Publishing. pp. 264 (2010)
- Stolc V, et al. *A gene expression map for the euchromatic genome of Drosophila melanogaster*. Science 306: 655–660 (2004)
- Stoletzki N and Eyre-Walker A. *Synonymous codon usage in Escherichia coli: selection for translational accuracy*. Mol Biol Evol. 24: 374–381 (2006)
- Stoletzki N. *Conflicting selection pressures on synonymous codon use in yeast suggest selection on mRNA secondary structures*. BMC Evol Biol. 8: 224 (2008)

- Stoletzki N and Eyre-Walker A. *Estimation of the neutrality index*. Mol Biol Evol. 28(1):63–70 (2011)
- Stone EA. *Joint genotyping on the fly: identifying variation among a sequenced panel of inbred lines*. Genome Res. 22(5): 966–974 (2012)
- Storz G. *An expanding universe of noncoding RNAs*. Science 296: 1260–1263 (2002)
- Strasburg JL, et al. *Genomic patterns of adaptive divergence between chromosomally differentiated sunflower species*. Mol Biol Evol. 26: 1341–1355 (2009)
- Strasburg JL, et al. *Effective population size is positively correlated with levels of adaptive divergence among annual sunflowers*. Mol Biol Evol. 28: 1569–1580 (2011)
- Stump AD, et al. *Centromere-proximal differentiation and speciation in Anopheles gambiae*. Proc Natl Acad Sci. USA 102: 15930–15935 (2005)
- Sturtevant AH. *Genetic studies on Drosophila simulans. I. Introduction. Hybrids with Drosophila melanogaster*. Genetics 5(5): 488–500 (1920)
- Sung W, et al. *Drift-barrier hypothesis and mutation-rate evolution*. Proc Natl Acad Sci. USA 109: 18488–18492 (2012)
- Sunyaev SR, et al. *SNP frequencies in human genes an excess of rare alleles and differing modes of selection*. Trends Genet. 16:335–337 (2000)
- Tachida, H. *Molecular evolution in a multisite nearly neutral model*. J Mol Evol. 50: 69–81 (2000)
- Tajima F. *Evolutionary relationship of DNA sequences in finite populations*. Genetics 105: 437–460 (1983)
- Tajima F. *Statistical method for testing the neutral mutation hypothesis by DNA polymorphism*. Genetics 123(3): 585–595 (1989)
- Tajima F. *Simple methods for testing the molecular evolutionary clock hypothesis*. Genetics 135(2): 599–607 (1993)
- Takahashi A, et al. *Genetic variation versus recombination rate in a structured population of mice*. Mol Biol Evol. 21: 404–409 (2004)
- Tamura K. *Estimation of the number of nucleotide substitutions when there are strong transition-transversion and G+C content biases*. Mol Biol Evol. 9: 678–687 (1992)
- Tamura K, et al. *Temporal patterns of fruit fly (Drosophila) evolution revealed by mutation clocks*. Mol Biol Evol. 21(1): 36–44 (2004)
- Tenaillon MI, et al. *Patterns of DNA sequence polymorphism along chromosome 1 of maize (Zea mays ssp mays L.)*. Proc Natl Acad Sci. USA 98: 9161–9166 (2001)
- The C. elegans Sequencing Consortium. *Genome sequence of the nematode C. elegans: a platform for investigating biology*. Science 282(5396): 2012–2018 (1998)
- The ENCODE Project Consortium. *An integrated encyclopedia of DNA elements in the human genome*. Nature 489: 57–74 (2012)
- The modENCODE Consortium. *Identification of functional elements and regulatory circuits by Drosophila modENCODE*. Science 330: 1787–1797 (2010)
- Thornton KR and Andolfatto P. *Approximate Bayesian inference reveals evidence for a recent, severe bottleneck in a Netherlands population of Drosophila melanogaster*. Genetics 172(3): 1607–1619 (2006)

- Thornton KR, et al. *X chromosomes and autosomes evolve at similar rates in Drosophila: no evidence for faster-X protein evolution*. *Genome Res.* 16:498–504 (2006)
- Thornton KR, et al. *Progress and prospects in mapping recent selection in the genome*. *Heredity* 98: 340–348 (2007)
- Tishkoff SA, et al. *Convergent adaptation of human lactase persistence in Africa and Europe*. *Nat. Genet.* 39(1): 31–40 (2007)
- Torgerson DG, et al. *Evolutionary processes acting on candidate cis-regulatory regions in humans inferred from patterns of polymorphism and divergence*. *PLoS Genet.* 5: e1000592 (2009)
- True JR, et al. *Differences in crossover frequency and distribution among three sibling species of Drosophila*. *Genetics* 142: 507–23 (1996)
- Tschermak V. *Über künstliche Kreuzung von Pisum sativum*. *Z. landwirtsch. Versuchsw. in Österreich*. 3: 465–555 (1900)
- Tupy JL, et al. *Identification of putative noncoding polyadenylated transcripts in Drosophila melanogaster*. *Proc Natl Acad Sci. USA* 102(15): 5495–5500 (2005)
- Venter JC, et al. *The sequence of the human genome*. *Science* 291: 1304–1351 (2001)
- Vicoso B and Charlesworth B. *Evolution on the X chromosome: unusual patterns and processes*. *Nat Rev Genet.* 7: 645–653 (2006)
- Vicoso B and Charlesworth B. *Effective population size and the Faster-X effect: an extended model*. *Evolution* 63: 2413–2426 (2009)
- Vitalis R, et al. *Interpretation of variation across marker loci as evidence of selection*. *Genetics* 158(4): 1811–1823 (2001)
- Vitti JJ, et al. *Detecting natural selection in genomic data*. *Annu Rev Genet.* 47: 97–120 (2013)
- Voight BF, et al. *A map of recent positive selection in the human genome*. *PLoS Biol.* 4: e72 (2006)
- Vries H. *Das Spaltungsgesetz der Bastarde*. Berichte der Deutsche Botanischen Gesellschaft 18: 83–90 (1900)
- Wakeley J. *Coevalent Theory. An Introduction*. Greenwood Village: Roberts and Company Publishers (2009)
- Wang ET, et al. *Global landscape of recent inferred Darwinian selection for Homo sapiens*. *Proc Natl Acad Sci. USA* 103(1): 135–140 (2006)
- Wang J and Caballero A. *Developments in predicting the effective size of subdivided populations*. *Heredity* 82: 212–226 (1999)
- Warnecke T and Hurst LD. *Evidence for a trade-off between translational efficiency and splicing regulation in determining synonymous codon usage in Drosophila melanogaster*. *Mol Biol Evol*. 24: 2755–2762 (2007)
- Watterson GA. *On the number of segregating sites in genetical models without recombination*. *Theor Popul Biol.* 7: 256–276 (1975)
- Watterson GA. *The homozygosity test of neutrality*. *Genetics* 88(2): 405–17 (1978)
- Watterson GA. *Mutant substitutions at linked nucleotide sites*. *Adv Appl Probab.* 14: 206–224 (1982)
- Watterson GA. *Allele frequencies after a bottleneck*. *Theor Popul Biol.* 26: 387–407 (1984)

- Weinberg W. *Über den Nachweis der Vererbung beim Menschen*. Jahreshfte des Vereins für vaterländische Naturkunde in Württemberg 64: 368–382 (1908)
- Weinreich DM and Rand DM. *Contrasting patterns of nonneutral evolution in proteins encoded in nuclear and mitochondrial genomes*. Genetics 156: 385–399 (2000)
- Welch JJ. *Estimating the genome-wide rate of adaptive protein evolution in Drosophila*. Genetics 173: 821–837 (2006)
- Wickham H. *ggplot2: Elegant Graphics for Data Analysis*. Springer New York, New York (2009)
- Wilson AC, et al. *Biochemical Evolution*. Annu Rev Biochemistry 46: 573–639 (1977)
- Wilson DJ, et al. *A population genetics-phylogenetics approach to inferring natural selection in coding sequences*. PLoS Genet. 7: e1002395 (2011)
- Wiehe TH and Stephan W. *Analysis of a genetic hitchhiking model, and its application to DNA polymorphism data from Drosophila melanogaster*. Mol Biol Evol. 10: 842–854 (1993)
- Wray GA, et al. *The evolution of transcriptional regulation in eukaryotes*. Mol Biol Evol. 20: 1377–1419 (2003)
- Wray GA. *The evolutionary significance of cis-regulatory mutations*. Nat Rev Genet. 8: 206–216 (2007)
- Wright S. *The evolution of dominance. Comment on Dr. Fisher's reply*. Am Nat. 63: 556–561 (1929)
- Wright S. *Evolution in Mendelian Populations*. Genetics 16: 97–159 (1931)
- Wright S. *Physiological and evolutionary theories of dominance*. Am Nat. 68: 25–53 (1934)
- Wright S. *The distribution of gene frequencies in populations*. Proc Natl Acad Sci. USA 23: 307–320 (1937)
- Wright S. *The distribution of gene frequencies under irreversible mutation*. Proc Natl Acad Sci. USA 24(7): 253–259 (1938a)
- Wright S. *Size of population and breeding structure in relation to evolution*. Science 87: 430–431 (1938b)
- Wright SI and Charlesworth B. *The HKA test revisited*. Genetics 168(2): 1071–1076 (2004)
- Xia Q, et al. *Complete resequencing of 40 genomes reveals domestication events and genes in silkworm (Bombyx)*. Science 326: 433–436 (2009)
- Yang Z. *PAML: a program package for phylogenetic analysis by maximum likelihood*. Comput Appl Biosci. 13: 555–556 (1997)
- Zeng K and Charlesworth B. *Studying patterns of recent evolution at synonymous sites and intronic sites in Drosophila melanogaster*. J Mol Evol. 70: 116–128 (2010)
- Zhang C, et al. *A whole genome long-range haplotype (WGLRH) test for detecting imprints of positive selection in human populations*. Bioinformatics 22(17): 2122–2128 (2006)
- Zhang L and Li W. *Human SNPs reveal no evidence of frequent positive selection*. Mol Biol Evol. 22: 2504–2507 (2005)
- Zhen Y and Andolfatto P. *Methods to detect selection on noncoding DNA*. Methods Mol Biol. 856: 141–159 (2012)

Zhu L, et al. *Patterns of exon-intron architecture variation of genes in eukaryotic genomes*. BMC Genomics 10: 47 (2009)

Zuckerkandl E and Pauling L. *Evolutionary Divergence and Convergence in Proteins*. Academic Press (1965)

Zuk O, et al. *The mystery of missing heritability: Genetic interactions create phantom heritability*. Proc Natl Acad Sci. USA 109: 1193–1198 (2012)

## **ANEXO**

---

Por motivos de espacio el anexo (junto con el resto de la tesis) en formato digital se pueden descargar en el siguiente enlace:

<https://github.com/CastellanoED/Thesis.git>

Del mismo modo en la contraportada se puede enlazar mediante un código QR a dichos contenidos





