A Critical Assessment of Storytelling: Gene Ontology Categories and the Importance of Validating Genomic Scans

Pavlos Pavlidis,*[1] Jeffrey D. Jensen,[2] Wolfgang Stephan,[3] and Alexandros Stamatakis[1]

[1]The Exelixis Lab, Scientific Computing Group, Heidelberg Institute for Theoretical Studies (HITS gGmbH), Heidelberg, Germany

[2]Ecole Polytechnique Fédérale de Lausanne, School of Life Sciences, Lausanne, Switzerland

[3]Section of Evolutionary Biology, Biocenter, University of Munich, Planegg-Martinsried, Germany

***Corresponding author:** E-mail: pavlidisp@gmail.com.

Associate editor: Arndt von Haeseler

Abstract

In the age of whole-genome population genetics, so-called genomic scan studies often conclude with a long list of putatively selected loci. These lists are then further scrutinized to annotate these regions by gene function, corresponding biological processes, expression levels, or gene networks. Such annotations are often used to assess and/or verify the validity of the genome scan and the statistical methods that have been used to perform the analyses. Furthermore, these results are frequently considered to validate "true-positives" if the identified regions make biological sense a posteriori. Here, we show that this approach can be potentially misleading. By simulating neutral evolutionary histories, we demonstrate that it is possible not only to obtain an extremely high false-positive rate but also to make biological sense out of the false-positives and construct a sensible biological narrative. Results are compared with a recent polymorphism data set from *Drosophila melanogaster*.

Key words: genome scanning, positive selection, gene ontology, validation, literature mining.

Introduction

An important problem in evolutionary biology is the identification of genes that have undergone recent positive selection. For the *Drosophila melanogaster* lineage for instance, the search for positively selected genes has generated a rapidly growing list of candidates. A multitude of studies have examined cosmopolitan and ancestral African populations of *D. melanogaster*, with a particular focus on genes that are important for local adaptation (e.g., Jensen et al. 2007; Tauber et al. 2007; Svetec et al. 2011). For example, Beisswanger et al. (2006) found a selective sweep in an ancestral African population at *ph-p* that may have led to the divergence of the *polyhomeotic* genes *ph-p* and *ph-d* (Beisswanger and Stephan 2008). Jensen et al. (2007) detected a selective sweep in the proximity of the *diminutive* gene for populations from China and Zimbabwe, consistent with the role of *diminutive* as a body size regulator and the observed clinal pattern variation of the body size trait. Tauber et al. (2007) reported that a mutation in the circadian clock gene *timeless* has spread in the European populations of *D. melanogaster*, favored by natural selection. The *timeless* gene affects diapause in response to light and temperature changes. Analyzing X-linked quantitative trait loci (QTL) affecting cold tolerance, Svetec et al. (2011) found evidence for a selective sweep in the gene *CG16700* for a European population of *D. melanogaster*. The expression variation at the *CG16700* has been associated with cold tolerance in an American population of *D. melanogaster* (Ayroles et al. 2009).

In general, after conducting a genomic scan for positive selection, researchers focus on the functions and biological properties of the identified gene regions. The primary goal of such studies is to corroborate/verify the biological importance of the genes that have been identified as being positively selected. One secondary goal of such studies is to increase the credibility of the positive selection model. In other words, if a well described, reasonable scenario/narrative can be constructed from the identified genes, the credibility of the positive selection hypothesis is assumed to increase. Often, the computational and statistical approaches that were deployed for detecting positive selection are considered to be valid and/or correctly implemented because results are biologically sensible. Reasonable findings (such as *diminutive* or *timeless*) are treated as true-positive cases and, albeit indirectly, are perceived to yield some validity to the methods that were used to detect them.

Here, we argue that the approach of perceiving an increased validity of results, simply because they make sense, may be misleading in many cases. Our main argument is that the majority of the genes in a genome have important biological functions with respect to the development, physiology, or evolution of any organism under study. Even for genes with obscure roles, it may not represent a grand challenge to unravel some connections with key genes using literature mining tools such as iHOP (information Hyperlinked Over Proteins, Hoffmann and Valencia 2004, 2005) or pathway knowledge bases such as Reactome (Vastrik et al. 2007).

To this end, we assess whether scanning a genome for positive selection can result in the detection of "meaningful" genes, even if we know a priori that the genome is neutral. Therefore, we initially simulate a sample of neutral genomes,

subsequently conduct a statistical scan that is followed by a computational annotation for function.

We do not intend to criticize computational and statistical approaches for detecting positive selection (e.g., see recent reviews of Nielsen 2005; Sella et al. 2009; Pool et al. 2010; Stephan 2010a, 2010b) nor to put into question the need for a thorough analysis and quest to understand and interpret the results. Our intention is to critically assess a certain tendency for attaching functional and/or evolutionary importance to findings obtained from genomic scans to verify the validity of the results. In other words, we are trying to post a note of caution with respect to the strong (over-) interpretation of results given the plethora of potential sources for analytical errors.

Materials and Methods

Our analysis consists of four steps:

1) We simulate samples drawn from a population evolving according to a neutral Wright-Fisher model. The demographic parameters of the population were inferred by analyzing a set of *D. melanogaster* X chromosomes sampled in the Netherlands (Li and Stephan 2006).
2) We map the gene annotation of the *D. melanogaster* X chromosome to the simulated data.
3) We conduct whole-genome scan for positive selection.
4) We carry out an enrichment analysis of the candidate targets that are under selection.

The complete procedure is outlined in figure 1.

Simulations

We simulated 100 data sets of 40 X chromosomes each, under a neutral demographic scenario, which can be described by a recent and deep bottleneck as well as an ancestral expansion. This demographic scenario has been inferred by analyzing the X chromosome of a European population of *D. melanogaster* sampled in the Netherlands (Li and Stephan 2006). In particular, the coalescent-based software *MaCS* (Markovian Coalescent Simulator) (Chen et al. 2009) was used to generate samples drawn from a population, which evolves under the neutral Wright-Fisher model. *MaCS* approximates genealogies on a recombining chromosome via a sequential Markov process. The tool models distant recombination events as being independent but preserves dependencies for recombination events that are located more closely to each other. Therefore, *MaCS* is more accurate than the sequential Markovian approximations of the coalescent (SMC; McVean and Cardin 2005), where a genealogy only depends on the preceding genealogy. Furthermore, *MaCS* execution times scale linearly with the size of the simulated region and thereby the tool allows for simulating the entire X chromosome of *D. melanogaster* (>20 Mb). We used the recombination rates along the X chromosome as inferred by Fiston-Lavier et al. (2010) for the simulations. Because recombination rates decrease toward either end of the chromosome, we analyzed the genomic region of the X chromosome between 3 and 18 Mb. Very small recombination rates might introduce biases into *MaCS* simulations because *MaCS* neglects dependencies between coalescent trees of distant genomic regions. Recombination rates are greater than 2.5×10^{-8} per base pair for the part of the X chromosome under study. The mutation rate for generating the simulated data sets has been set to the average mutation rate along the X chromosome ($\mu = 1.45 \times 10^{-9}$) as estimated using the average divergence from *D. simulans* in Li and Stephan (2006).

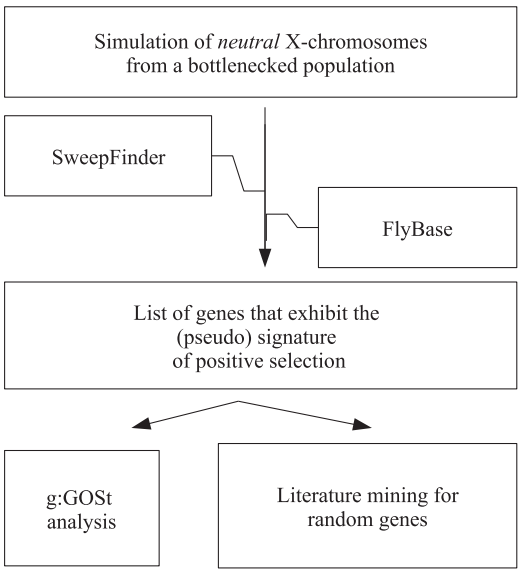

Fig. 1. Flowchart of the process for the identification and description of genes that exhibit pseudo-positive selection. The outlined process is repeated for each of the 100 simulated data sets.

Whole-Genome Scanning

We applied the widely used (e.g., Nielsen et al. 2005; Svetec et al. 2009; Pavlidis et al. 2010; Li et al. 2011) SweepFinder program (Nielsen et al. 2005) to detect genomic regions in which the site frequency spectrum (SFS) deviates from the neutral expectation (i.e., the average SFS over the entire genome) and resembles the SFS of a simple selective sweep model (see eq. 6 in Nielsen et al. 2005). SweepFinder is based on the composite likelihood ratio (CLR) test developed by Kim and Stephan (2002). It is, however, more robust to demographic assumptions because of using the average SFS over the entire genome as the expected neutral SFS. SweepFinder values represent a CLR between the maximum likelihood of a selective sweep model and a neutral model (represented by the average SFS of the genome). Thus, large values of SweepFinder (statistically significantly larger than neutrality) suggest a selective sweep. The CLR test of SweepFinder was applied to 10,000 equidistant points along the genome.

Identification of Putative Selection Regions

For each of the 100 simulated data sets, a cutoff value was calculated to identify regions that may exhibit positive selection. For each data set, the cutoff value was selected as the 99.5th percentile from the empirical distribution of the 10,000 equidistant points along the genome for which the CLR statistic of SweepFinder was calculated. We then mapped the

positions of those points exceeding the cutoff value (outliers) to gene names on the X chromosome using Flybase (version 5.35, March 2011). We only executed the mapping if the point was located within the gene or 2 kb upstream or downstream from the gene. We chose a setting of 2 kb because regulatory regions for gene expression are located in the proximity of genes. It has been reported that positive selection acts on regulatory regions (e.g., Torgenson et al. 2009; Cruickshank and Nista 2011). Furthermore, in genome scans for positive selection, genes that are located several kilobases away from the inferred target of selection are also reported as candidates (see e.g., Oleksyk et al. 2010 for a recent review). These genes represent the candidate loci for positive selection. If a single point was mapped to multiple (overlapping) genes then all genes were used in the analysis. Furthermore, if two or more points were mapped to the same gene, only the first point was used in the analysis. Points that could not be mapped to a gene using this procedure were excluded from the downstream analyses. We outline this process in figure 1. Figure 2 provides an example of this mapping process in the genomic region between 5 and 6 Mb for one of the simulated data sets. We study the properties of outlier points that were excluded from the analysis in supplementary section I, Supplementary Material online.

Enrichment Analysis of Candidate Gene Lists

The goal of the enrichment analysis of gene lists is to detect a set of concepts or terms that are enriched in the gene list relative to a reference (or background) gene list. We used the g:GOSt (Gene Ontology Statistics) module of g:Profiler (Reimand et al. 2007, 2011). g:GOSt performs statistical enrichment analysis to detect Gene Ontology terms, biological pathways, regulatory motifs, microRNAs, protein–protein interactions, or human disease annotations that are associated with the gene list at hand. Our main goal is to deploy the above machinery to search for biological annotations that are enriched in each candidate gene list.

Scanning a Real Data Set

In addition to the simulated data sets, we also scanned a real biological data set with SweepFinder. The data set comprises 37 inbred lines of *D. melanogaster* sampled in North Carolina and is publicly available (Drosophila Population Genetics Project [DPGP, http://www.dpgp.org]). This population of *D. melanogaster* might have experienced more complex demographic changes than the simulated data sets. For example, it has been reported that American populations of *D. melanogaster* are admixed with African alleles (Caracristi and Schlötterer 2003). It is widely accepted, however, that the demography of non-African populations of *D. melanogaster* is characterized by deep and very recent bottlenecks, similar to our simulated demographic model (e.g., model BN3 in Haddrill et al. 2005; Yukilevich et al. 2010). On this real data set, SweepFinder was executed using the same settings as for the simulated data. The enrichment analysis on the real data set was also conducted in exactly the same way as for the simulated data. We included a real data set to conduct a quantitative comparison of simulated and real data results such as to ensure that our results are quantitatively analogous and comparable with those observed in real

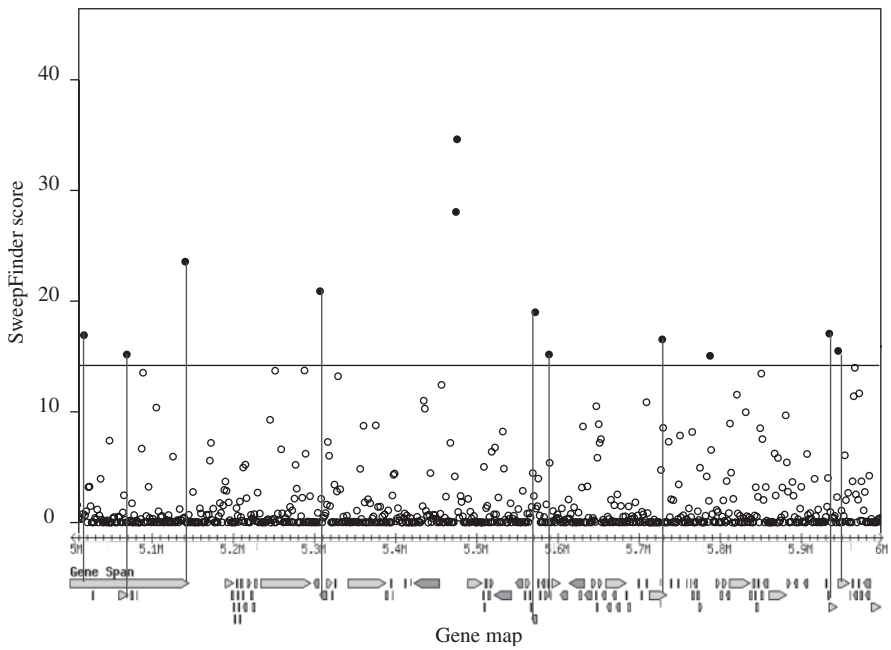

Fig. 2. Mapping of points exceeding the threshold value to genes in Flybase (version 5.35, March 2011). Vertical lines connect points to genes located at the corresponding positions. Points above the threshold line are shown as filled circles, whereas points below are shown as empty circles. Points above the threshold line that could not be mapped to a gene were discarded. Multiple hits to a certain gene were excluded from the analysis. The horizontal line depicts the 99.5% cutoff value defined from the empirical distribution of the SweepFinder scores along the genome. For illustrative purposes only the region between 5 and 6 Mb is shown. The genes that can be mapped to the points define the set of candidate genes for positive selection.

data. To verify this analogy, we compared the distributions of SweepFinder output values between simulated and real data sets, the distribution of SweepFinder outliers, as well as the number of overrepresented biological annotations as identified by the enrichment analysis.

Results

Initially, we summarize the results obtained on the 100 simulated data sets. For discussing specific issues, we always refer to the first simulated data set. Because the (arbitrary) order of the simulated replicates is independent of their content, we chose to provide specific examples from the first simulated data set for convenience, provided that it contained at least one significant biological term or property. The first simulated data set is not an outlier among the simulated data sets with respect to the number of significant biological categories or SweepFinder results (see Enrichment Analysis of Genes that Exhibit the Signature of Positive Selection in Simulations for a summary across all simulated data sets). The first simulated data set, as well as software/scripts that reproduces the analysis can be downloaded at http://exelixis-lab.org/pavlos/gopopgen.tar.gz.

Mapping Positive Selection and Performing Enrichment Analysis

Mapping Positive Selection in Simulated Data Sets

As described earlier, we generated a list of genes encompassed within significant regions for each data set. On average, 43 outlier genes (minimum 27 and maximum 60) were detected (see also supplementary section II, Supplementary Material online). Here, polymorphic patterns that exhibit the signature of a selective event are associated with a shift of the SFS toward low- and high-frequency–derived alleles. An example for such a polymorphic pattern is provided in figure 3 for the genomic region with the highest SweepFinder score in the first simulated data set. This polymorphic pattern resembles the expectation of a recent selective event, which is characterized by the lack of intermediate-frequency alleles. Along a recombining genome, such patterns occur due to the stochastic alternation of shallow and deep coalescent trees. Qualitatively, polymorphic patterns that are characterized by the lack of intermediate-frequency alleles and generated by the stochastic alternation of shallow and deep coalescents are very similar to patterns that have been generated by a "genuine" selective sweep. Therefore, it may be impossible or will be very difficult to distinguish them from true-positive selection events. Thus, results do not only reflect the sensitivity of the neutrality tests to demographic assumptions, but, more importantly, the resemblance of the polymorphic patterns created by some neutral and non-neutral evolutionary scenarios, as has been well described in previous studies (e.g., Barton 1998; Galtier et al. 2000; Przeworski 2002; Depaulis et al. 2003; Jensen et al. 2005; Nielsen et al. 2005; Li and Stephan 2006; Pavlidis et al. 2010).

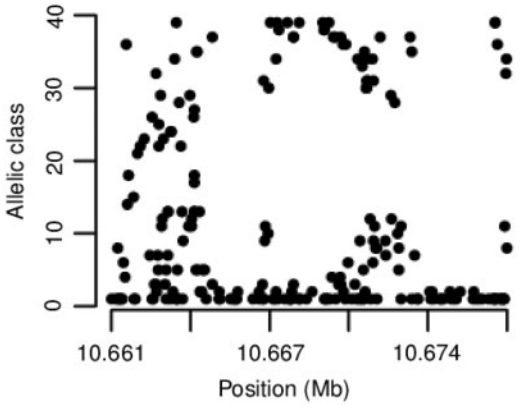

Fig. 3. Patterns of the SFS classes that occur around the highest SweepFinder peak in the first simulated data set. The graph denotes the class of the polymorphisms (*y*-axis) at different positions (*x*-axis). For example, a point at position 10,670,000 with allelic class 3 denotes that at this position there exists a tripleton in the data set. One of the characteristic signatures of a recent selective event, which is the lack of intermediate-frequency–derived polymorphisms, is illustrated here as a blank region that occurs in the center of the graph.

Enrichment Analysis of Genes that Exhibit a Signature of Positive Selection in Simulations

As already mentioned, enrichment analyses for gene ontology (GO) categories, pathways, transcription factor binding sites, microRNAs, diseases, and protein–protein interactions were performed using g:GOSt. g:GOSt uses a custom algorithm g:SCS (Set Counts and Sizes) for performing multiple testing correction over GO categories, pathways, and other types of functional data. For a given input gene query of fixed length, the correction analytically approximates the empirical 95% quantile of *P* values obtained from randomly generated gene lists of the same fixed length, separately and independently for each data source (e.g., GO, KEGG, Reactome). According to simulations by Reimand et al. (2007), SCS correction produces values that are comparable with two standard multiple testing procedures: being more conservative than the Benjamini-Hochberg false discovery rate (Benjamini and Hochberg 1995) and less conservative than the Bonferroni correction. To further test the SCS correction approach, we performed enrichment analysis for random subsets of 50 X chromosome genes. Among these, 34% yield at least one significant category and 9% yield at least two significant categories (see also supplementary section III, Supplementary Material online). g:GOSt returns scaled *P* values such as the highest significant *P* value after multiple testing correction assumes a value of 5% and smaller *P* values are scaled proportionally. The background data set used in the enrichment analysis entailed only genes located on the X chromosome, as downloaded from the Ensembl genome browser (Flicek et al. 2011).

In the first simulated data set, we find an enrichment for eight GO categories (six refer to biological process and two to cellular component) and three related to the Human Phenotype Ontology (HPO). This means that homologues

in humans were enriched in the respective HPO categories. Furthermore, we detected 1 miRNA, 2 Reactome and 1 transcription factor binding site categories. For instance, the *P* value of the GO category "cellular nitrogen compound biosynthetic process" is 4.16×10^{-3}, and for the "cell adhesion" category the *P* value is 2.99×10^{-4}. On average, 5.19 statistically significant categories were detected per simulated data set (a total of 519 significant categories in 100 simulated data sets).

Out of 100 simulated data sets, 77 yielded at least one and 64 at least two significant categories: 13, 11, and 10 data sets contained one, two, and three statistical significant categories, respectively and 16 data sets gave rise to more than 10 significant categories. The maximum number of significant categories detected in a single simulated data set was 25 (45 genes were outliers). Categories in enrichment analysis are not independent of each other since a set of genes may form part of several categories and genes can also form hierarchies. Thus, detecting 25 categories using 45 genes does not represent an unexpected result. A description of significant functional categories as inferred for simulated data sets is provided in supplementary section IV, Supplementary Material online.

Enrichment Analysis of the Genes that Exhibit a Signature of Positive Selection in the DPGP Data Set

We performed enrichment analysis of the DPGP data set in analogy to the simulated data sets. The command line parameters and program invocations used to detect the candidate genomic regions for positive selection and the enrichment analysis were exactly identical for the DPGP data set and the simulated data sets. Here, the goal consists in comparing the number of statistically significant biological terms detected in a real and 100 neutrally evolving simulated data sets. Therefore, our comparison is not qualitative but quantitative. Furthermore, we do not intend to argue that the DPGP data set has necessarily evolved neutrally, but rather, that it may be extremely difficult to accurately determine whether a data set has evolved under positive selection or under neutrality by mining biological meta-information.

Using g:GOSt again, nine statistically significant biological terms were obtained for the real data set related to transcription factor binding sites. *P* values ranged from 4.82×10^{-2} to 3.26×10^{-4}. Thus, the number of biological terms obtained using g:GOSt is not higher in the real data set than in the simulated data sets (remember that on average 5.19 terms are detected per data set and 16 of 100 simulated data sets contained at least 10 biological terms).

Believing the Results from Simulations—Is It Possible to Construct Plausible Evolutionary Explanations from Pseudo-Selection Results?

Here, we intentionally treat the first simulated neutral data set as real biological data sampled from a European population of *D. melanogaster*. Consequently, the SweepFinder results comprise a candidate list of genes for positive selection, and we try to explicitly construct a narrative for the results. We examine whether it is possible to mine the literature to provide further evidence in support of the adaptive role of the artificial candidate genes during the colonization of Europe by *D. melanogaster*. To this end, we chose the three genes with the highest SweepFinder values from the first simulated data set and interpret them as true-positive results.

Narrative for Gene CG15211

The first gene is referred in Flybase by the symbol Dmel\CG15211. Despite the limited amount of information for *CG15211*, there exist indications that it is involved in important biological processes. The GO indicates that its molecular function and the biological process it is involved in are unknown. In addition, no phenotypic data are available. However, there exist four annotated transcripts and four annotated polypeptides, and it also contains a MARVEL-like domain (see Flybase, http://flybase.org/reports/FBgn0030234.html for further information). All described proteins containing the MARVEL domain are consistent with the M-shaped topology; that is, they contain four transmembrane helix regions. Their function could be related to cholesterol-rich membrane apposition events (see also http://www.ebi.ac.uk/interpro/IEntry?ac=IPR021128).

Using a statistically rigorous analysis for identifying genes whose expression was altered during starvation stress treatment, Harbison (2004) showed that *CG15211* is down-regulated in starvation-stress experiments. Furthermore, Wijnen et al. (2006) demonstrated that *CG15211* transcript levels are light-regulated. Thus, *CG15211* may have played an important role during the initial stages of European colonization following the migration of the founding population from Africa. Because of daylight and habitat differences between African and European environments, the first populations that colonized Europe may have experienced a high selective pressure on the *CG15211* gene.

Narrative for Gene CG8188

The second identified gene is the Dmel/CG8188. It is a protein coding gene and it is predicted (based on sequence similarity) to exhibit ubiquitin–protein ligase activity. According to the GO annotation, the biological processes in which it is involved are unknown. Eleven alleles have been reported and no phenotypic data are available. *CG8188* appears to affect the wing formation particularly in males. *P*-element insertions in *CG8188* caused general and localized deformations of the wings of males and only general deformations of the wings of females (Carreira et al. 2011). Its human homologue, UBE2S/E2-EPF, drives elongation of ubiquitin chains by the anaphase-promoting complex (Wu et al. 2010). E2-EPF has been reported to be highly expressed in common human cancers (Ohh 2006). According to BioGRID (Stark et al. 2011), CG8188 interacts with seven proteins encoded by *CG5053, TRAF6, AOP, DM (diminutive), LOCO, ORG-1*, and *TIM (timeless)*. Interactions have been demonstrated experimentally, either with two-hybrid systems or with affinity capture-western blots. *CG5053* is important for the compound eye development (inferred from a mutant phenotype, Flybase; Langton et al. 2009). *AOP* participates in many important functions from the compound eye development

(Flybase; Li and Carthew 2005) to gonad, muscle, and neuron development as well as apoptosis. TRAF6 is associated with defense response to Gram-negative bacteria and innate immune response (Flybase; Leulier et al. 2006). It has been reported that *TIM* (Tauber et al. 2007) and *DM* (Jensen et al. 2007) have experienced recent selective sweeps in non-African populations. For the *CG8188* gene, we were not able to find direct support for a selective sweep. However, we demonstrated that *CG8188* can interact with important developmental genes (*CG5053* and *AOP*) or with genes for which there exists evidence for selective sweeps (*DM* and *TIM*). Thus, *CG8188* may have played a significant role during the adaptation of the European *D. melanogaster* population via its interaction with functionally important genes.

Narrative for Gene CG6788

The third gene is the Dmel/CG6788. Its molecular function is predicted to be receptor binding based on InterPro domains and it is suggested (based on sequence similarity) to be involved in cell adhesion and it also contains a fibrinogen domain. Further details can be found at Flybase (http://flybase.org/reports/FBgn0030880.html). Arbeitman et al. (2004) showed that the *CG6788* gene is expressed only during the 0–24 h stage of adult male life in the male ejaculatory bulb. According to Arbeitman et al. (2004) the timing of *CG6788* expression points toward a role in ejaculatory bulb development. Furthermore, differential expression analyses revealed that *CG6788* participates in immune response. In a microarray analysis, Chamilos et al. (2008) showed that *CG6788* is one of 98 genes whose induction was upregulated after infection with *Rhizopus* compared with the control aseptic injury or those injected with *A. fumigatus*. There also exists evidence that *CG6788* is activated by Toll. Members of the Toll receptor family play an important role in activating immunity genes because the Toll pathway controls the induction of antimicrobial peptide genes. *CG6788* was found to be highly activated (~44 times) in the constitutively active *Toll* mutant larvae (Bettencourt et al. 2004). Phenotypic plasticity may be related with the expression levels of *CG6788*. Sambandan et al. (2008) showed that *CG6788* variation in expression due to the larval growth medium depends on the genotype. Thus, *CG6788* appears to be an important gene for immune response. Its expression level is regulated by Toll, and appears to have significant effects on the genotype by environment interactions; its expression levels change due to the larval growth medium, but the degree of change also depends on the environment. Positive selection may have acted on the gene or its *cis*-regulatory sequences after the colonization of Europe to contribute to the adaptation of the newly founded population in the European environment.

Impact of Demography

We show that meaningful narratives can be constructed from neutral simulations by using an inferred demographic model for *D. melanogaster* sampled in the Netherlands (Li and Stephan 2006). This demographic scenario is characterized by a deep and recent bottleneck scenario. It has been shown (Pavlidis et al. 2010) that severe bottlenecks considerably increase the proportion of false-positives in neutrality tests. Even if demography affects our analysis, the results are not limited to the specific demographic model because 1) milder bottlenecks (e.g., Thornton and Andolfatto 2006) can also generate a relatively high proportion of false-positives (e.g., true-positives: 0.40–0.67; false-positives: 0.22–0.35; table 4, Pavlidis et al. 2010), and 2) many natural populations have experienced recent founder events, where the number of founders ranges from a few to a few hundred individuals (Pascual et al. 2007). Therefore, severe bottlenecks (as used here) represent a realistic scenario for natural populations. Finally, genome scans for positive selection detect candidates as outliers in the distribution of a test statistic (e.g., Kelley et al. 2006; Gu et al. 2009; Rubin et al. 2010). The actual demography of the natural population affects the proportion of true-positives at the tail of the distribution because false-positives form a part of the tail. Using simulations, we demonstrate that even if true-positives are completely absent from the tail of the distribution of a test statistic, it is still possible to come up with convincing narratives.

We simulated 80 X chromosome data sets using a milder bottleneck model (Thornton and Andolfatto 2006) as opposed to Li and Stephan (2006) to further generalize the results. The average minimum, median, and maximum SweepFinder values and the 95% confidence intervals for the milder model are 0 (0–0), 0.12 (0.09, 0.15), 42.3 (33.07, 53.74), respectively. The corresponding average minimum, median, and maximum values for the Li and Stephan (2006) model and the 95% confidence intervals are 0 (0–0), 0.21 (0.18–0.23), and 56.4 (42.8–81.9). Simulations involving even milder bottlenecks than those used above (that have approximately the same severity) further reduce the CLR values of SweepFinder (PP, unpublished results). Applying g:GOSt to the simulated data with the Thornton and Andolfatto (2006) milder bottleneck model using exactly identical parameters as for the deep bottleneck model, still yields significant categories in 85% of the data sets. The maximum number of significant terms for a single data set was 74 (16 of them refer to GO terms and the remaining 58 to transcription factor binding sites).

Testing the Uniformity of Outliers in a Whole-Genome Scan for Selective Sweeps

One may argue that instead of performing whole-genome neutral simulations to test the discovery of statistically significant functional categories, we could uniformly place 50 (=0.005 × 10,000) points on the genome and assume that they are outliers of a hypothetical selective sweep scan. Such an approach is biased because SweepFinder outliers are not uniformly distributed. Depending on the recombination rate and the demographic model, clusters of outliers instead of uniformly distributed points are generated. Clustering of outliers might be pronounced for some demographic histories (such as bottlenecks) and variable recombination rate along the chromosome (as we have assumed in this study).

Furthermore, bottlenecks may result in very different coalescent trees along a recombining chromosome. Consequently, regions with low variation, a shifted site frequency spectra or high linkage disequilibria may be generated locally on the chromosome, and thus producing patterns of pseudo-selective sweeps. The discovery of statistically significant GO categories (see also Discussion) can thus be attributed to this local generation of spurious selective sweep patterns. Recombination also plays an important role in the generation of pseudo-selective sweep patterns because coalescent histories may be different across a recombination breakpoint. Too large recombination rates will eventually result in a different genealogy every few (or even a single) base pairs, thereby blurring any selective sweep pattern. Absence of recombination, on the other hand, will result in exactly the same underlying coalescent tree for the whole chromosome. Thus, regions that resemble a selective sweep cannot be produced. There is, however, a range of recombination rates that can produce spurious selective sweep patterns depending on the demographic scenario.

To test the hypothesis that recombination rate and demographic history are responsible for the clustering of outliers, we examine the uniformity of outliers under the following assumptions: 1) Li and Stephan's (2006) bottleneck model and recombination rates inferred by Fiston-Lavier et al. (2010), 2) Li and Stephan's (2006) bottleneck model using a 10 times lower recombination rate than the inferred by Fiston-Lavier et al. (2010), 3) same as used in 1) but recombination rate is 10 times higher, 4) standard neutral model with the same recombination rate as in 1), and 5) standard neutral model with 10 times lower recombination rate than in 1). For each of the five scenarios, we carried out 100 simulations. Uniformity of outliers was tested using the Kolmogorov-Smirnov test (KS) as implemented in R at a significance threshold of 0.05. More specifically, we assess the number of data sets that deviate from uniformity (KS P value < 0.05) in each of the five scenarios and compare it with the expected number of deviations under a uniform distribution. For scenario 1 the P value of 62 (5 are expected by chance) replicates is smaller than 0.05. For simulation 2, 99 data sets have a P value of < 0.05. On the other hand, the P value is smaller than 0.05 in only 9 data sets under scenario 3. For the standard neutral simulations, under scenario 4, we observe that 12 data sets show a P value of less than 0.05; and under scenario 5, 43 data sets have a P value of <0.05. Thus, uniformity increases under a standard neutral model (as opposed to a bottleneck model) and higher recombination rates, and consequently the joint action of demography and recombination, affects the clustering of outliers. The P value distributions under all five simulation scenarios are described in supplementary section V, Supplementary Material online.

The results of the enrichment analysis are not substantially affected by the recombination rate. For the scenario 2, where recombination is 10 times lower, 69 data sets and 43 data sets have at least one and two significant categories, respectively. When recombination is 10 times higher (scenario 3), 81 data sets show at least one significant category and 70 data sets at least two significant categories.

Using the ω Statistic and the Combination of the ω Statistic and SweepFinder to Detect Selective Sweeps

The ω statistic (Kim and Nielsen 2004; Jensen et al. 2007; Pavlidis et al. 2010) uses linkage disequilibrium (LD) patterns to detect recent positive selection. As the ω statistic and SweepFinder detect different signatures of selective sweeps (LD vs. SFS), the two approaches can be used in a complementary way. Furthermore, combining the ω statistic and SweepFinder using a machine learning technique increases the power to detect selection (Pavlidis et al. 2010). Here, we used a more recent highly optimized code for computing the ω statistic, OmegaPlus (http://www.exelixis-lab.org/software.html). We tested two analysis options: 1) applying the ω statistic assuming the 99.5th percentile as threshold (similarly to the SweepFinder analysis), and 2) combining results from SweepFinder and the ω statistic assuming 50 points to be outliers (0.5%) that are entailed in the top-right rectangle of the scatter-plot between SweepFinder and the ω statistic (supplementary section VI, Supplementary Material online).

The distribution of outliers along the X chromosome is different between the ω statistic and SweepFinder. ω Statistic outliers are distributed more uniformly than SweepFinder outliers: only in three data sets does the KS-test yield P values of less than 5% (vs. 62 data sets in SweepFinder). For the combination analysis using the ω statistic with SweepFinder, 21 data sets deviate from uniformity. The difference of the outlier distributions as obtained by the ω statistic and SweepFinder can be attributed to the distribution of coalescent trees along the genome (supplementary section VII, Supplementary Material online). Enrichment analysis based on the ω statistic yields 83 data sets with at least one significant category and 62 data sets with at least two significant categories. Results from the joint analysis of the ω statistic and SweepFinder are similar: 79 and 62 data sets yield at least one and two significant categories, respectively.

Analysis with the ω statistic suggests that even if the distribution of outliers would be uniform the enrichment analysis is still biased. Apparently, this is because longer genes are favored to be in the list of candidate genes when outliers are distributed uniformly. Consequently, gene length will affect the enrichment analyses if there exist categories with a preponderance for either short or long genes. This bias has been recently addressed in RNA-seq studies (Young et al. 2010). Thus, even when analyses are performed with tools that alleviate the problem of outlier uniformity (e.g., the ω statistic), current implementations for enrichment analyses are inappropriate for whole-genome scans of positive selection.

Discussion

Why Do We Believe Results if They Make Sense?

Consider an abstract genome scan study that yields a list of candidate genes for positive selection, and let A be the associations of these genes to biological meta-information such as

functions, processes they are involved in, and expression regulation. Let M represent a model of positive selection with probability $P(M)$. The likelihood of M given the association data A is $P(A|M)$. Then, *intuitively*, we believe that $P(A|M)$ should be larger than $P(A)$; that is, the probability to find meaningful biological descriptions of the data should be larger under the model of positive selection than the total probability of A $P(A)$. Thus, according to Bayes' rule $P(M|A) = P(M) \times P(A|M)/P(A)$, our posterior probability $P(M|A)$ for the model M is increased when biological information about the findings can be obtained. In contrast to this, we argue by means of a simple simulation study that $P(A)$ does actually not need to be smaller than $P(A|M)$.

Consider the following simple example that elucidates the above argument: assume that we are scanning a genome from a European *D. melanogaster* sample for positive selection. It will be *reasonable* and *intuitive* to expect that the presence of selection may affect genes that are responsible for cold-tolerance, starvation, pigmentation, light-cycle regulation, and resistance to insecticides. In general, we are biased toward searching for, or expecting to observe gene functions that are associated with the environmental differences between Europe (derived population) and Africa (ancestral population). In this example, M represents a model of positive selection, and A represents the associations with biological information that we believe to make sense. Therefore, $P(A|M)$, that is, the likelihood of selection, is expected to be high when A makes sense. Thus, if we can provide evidence that the candidate genes are indeed associated with environment-specific functions (or generally with reasonable functions in the context of the study), we automatically increase the likelihood of selection. The underlying problem is that we cannot estimate the total probability of A under selection or neutrality, $P(A)$. Here, we show that we can also effectively mine the literature to obtain biologically interesting annotations under the null (neutral) model. Thus, $P(A)$ does actually not need to be smaller than $P(A|M)$.

Do Neutrality Tests Detect Genomic Regions that Have Experienced Positive Selection?

Genomic scans for positive selection will detect regions with unusual polymorphic patterns, generated by unusual genealogies. If an empirical genome-wide distribution of a summary statistic is used to define a significance threshold, then these genomic regions are located at the respective tails of the distribution. A disadvantage of using empirical distributions for detecting regions that have experienced positive selection is that the proportion of the genome that has experienced positive selection is unknown, and therefore the number of detected genetic regions is likely to be over- or underestimated. Moreover, it is unknown whether true-positives will preferentially populate the tail of the distribution, particularly in nonequilibrium populations (Thornton and Jensen 2007; Pavlidis et al. 2010). On the other hand, if a parametric bootstrap method is used to define a cutoff value for a neutrality test, we may fail to precisely detect those regions that have experienced positive selection because of a plethora of simplifying assumptions that are inherent to the model. Thus, neither model-based nor empirical-distribution approaches can accurately identify genomic regions that are under positive selection mainly because they represent an incorrect null hypothesis (Thornton and Jensen 2007). Furthermore, both approaches will typically yield a list of candidate loci for positive selection. The proportion of true-positives in such a list is unknown, and GO categorization is often used to identify likely candidates.

Should We Believe Genome Scan Results Because They Make Sense?

A null hypothesis for what "makes sense" does not exist. Therefore, the validity of the results is independent of GO descriptions. Furthermore, we are unaware of the selective pressures that the organisms have experienced in the past and therefore can not expect a certain category of genes to be under positive selection. Even if we expect and accept some plausible scenarios such as cold tolerance of *D. melanogaster* in cold climates, the selective pressures that this specific environmental condition induces on the organisms under study and the various ways in which an organism may respond to this pressure are unknown. Finally, due to the pleiotropy of many genes and their complex relationships, we may only have limited knowledge on how a certain response actually manifested itself in nature.

Evidently, we can retrieve biological functions that are potentially related to adaptation for a significant number of genes. This is especially true when there is no a priori hypothesis to test for, or when the hypothesis at hand is very general. For example, the attempt to explain adaptation of *D. melanogaster* in Europe can involve a very large number of genes ranging from cold-tolerance, metabolism, immunity, learning to almost any other process or function that we consider as being important. However, it may be possible that adaptation affects only a very limited number of genes.

More importantly, the majority of genes can be put into context and construct a narrative that "makes sense," for instance when genes are linked to functionally important genes. Additionally, our imagination, ability, and desire to detect patterns, as well as the wealth of information available in the literature may lead us to come up with an explanation for the action of positive selection on an arbitrary gene in the genome.

It is worth noting that the overlap between lists of candidate genes for positive selection from various studies in the human populations is limited (e.g., Nielsen et al. 2005; Bustamante et al. 2005; Voight et al. 2006; Williamson et al. 2007; see also Oleksyk et al. 2010 for a recent review). In other words, candidate gene sets are mostly disjoint among studies. This lack of overlap might reflect the differences between the applied neutrality tests or that the neutrality tests cannot accurately detect the target of selection. Yet, GO enrichment analysis and literature mining for certain genes are extensively used in such studies not only to explain the findings but also to interpret the results as being meaningful.

Comparison between Observation and Simulations

Qualitatively, there are not substantial differences between the real reference data set we used and the 100 simulated *neutral* data sets. In the real data set, SweepFinder detects stronger selection (supplementary section VIII, Supplementary Material online). This might be an indication of true-selective sweeps in the real data set, but it can also suggest that the recombination rate or mutation rate used in the simulations was higher than in reality There is no evidence that SweepFinder CLR values are significantly higher in the real data set than in the simulated data sets for almost any quantile value (see also supplementary section VIII, Supplementary Material online). Of course, this may just mean that the demographic scenario used here is too conservative or that the power of SweepFinder to detect selection under the specific demographic scenario (Li and Stephan 2006) is too small. Furthermore, there is no evidence that the real data analysis yields more biological categories. This points into the direction that by almost randomly sampling genes across a genome, as was done in our simulations, we can construct equally good narratives as for the real data set. This is particularly worrisome because i) a priori we have no reason to believe that a certain proportion of the genome should have experienced a recent sweep, and ii) the tail of the distribution of neutrality tests is not even necessarily enriched in true-positives particularly in nonequilibrium populations (Thornton and Jensen 2007; Pavlidis et al. 2010).

Why Can Significant GO Categories Occur in Simulations?

Over-representation analyses returned many significant categories in neutral data sets. This is unexpected: random subsets of X chromosome genes result in significantly less hits (0.5 per random subset of 50 genes vs. 5.19 per simulation) on average (Mann Whitney test P value: 5.936×10^{-16}). However, specifically for a genome scan analysis, the outcome may not be truly random. There is increasing evidence that genes are not distributed randomly across chromosomes. By analyzing five eukaryotic genomes, Lee and Sonnhammer (2003) found evidence for clusters of genes that belong to the same pathway. Al-Shahrour et al. (2010) show that in well-annotated species, genomes are formed by a large amount of functional neighborhoods. Approximately 5.3% of genes belong to functional neighborhoods in *D. melanogaster*, 3% of the genes belong to functional neighborhoods in *A. thaliana*, 7.2% in humans, and 12.8% in *R. norvegicus*. Neighboring genomic regions do not evolve independently. If the recombination rate is not very large, the coalescent histories of neighboring genomic regions are expected to be more congruent. Polymorphic patterns and thereby neutrality test scores are based on coalescent histories. Thus, neutrality tests might yield high values in large genomic regions because of dependent evolutionary histories in neighboring locations. Consequently, outliers will tend to form clusters. Combining functional clustering with evolutionary histories at genomic neighborhoods may thus eventually cause over-representation of some biological categories.

Using the ω statistic or combining SweepFinder with the ω statistic, the clustering of outliers is alleviated. However, several functional terms are detected as significant in enrichment analyses. This is because current implementations of enrichment analyses neglect important factors such as the length of genes.

A further evolutionary hypothesis that could lead to statistical enrichment of some biological categories is the combination of population bottlenecks with purifying selection. If purifying selection affects preferentially a specific set of genes (e.g., a metabolic pathway) and the population undergoes a bottleneck, then depletion of genomic variation and SFS skew will become more pronounced for the genes of the pathway. These genes could thus be detected by neutrality tests as candidates for positive selection. A subsequent GO enrichment analysis would then report the metabolic pathway under consideration.

Verifiability and Reproducibility of Results

A rarely addressed problem in genome analyses is verifiability and reproducibility of results. Often, arbitrary or hard to justify decisions are taken to determine and set data analysis parameters. Even small variations of these parameters can have dramatic effects on the final results. For example, arbitrary window size and offset values are used in scans for positive selection with simple summary statistics (e.g., Tajima's [1989] D). Small changes (even by only a few bases) of window size might change the results from being statistically significant to being nonsignificant (supplementary section IX, Supplementary Material online). Often, these (arbitrary) parameter settings are not reported in the literature, and consequently reproducing and verifying computational experiments becomes hard or even impossible. For example, Akey et al. (2002) considered as outliers the top 2.5% of an empirical distribution of F_{ST} values, and Kayser et al. (2003) assumed that candidate loci for positive selection exhibit unusual high R_{ST} and/or ln (RV) values. R_{ST} is analogous to F_{ST} and measures population differentiation. Large positive or negative values of ln (RV) might reflect a recent selective sweep at a nearby locus in a population (Kayser et al. 2003). In our analysis, we have also used some arbitrary settings for several quantities and thresholds that are typically required to conduct genome scans for positive selection. For example, we ignored the first 3 Mb of the chromosome because of small recombination rates and we used the 99.5th percentile as threshold in SweepFinder. Arbitrary or hard to justify parameter decisions are mostly inevitable in data analyses and they should be considered prior to examining the results. Another problematic phenomenon is the continuous data growth that yields the task of repeating computations for reproducing results nearly impossible because of excessive computational requirements (Stamatakis and Izquierdo-Carrasco 2011). Software bugs, nondeterminism in parallel codes, and the usage of different code versions may also produce unverifiable results. The constant changes in computer architectures, compilers, and scientific libraries further complicate the reproducibility of experiments. For

example, in the current analysis, *MaCS* (v.0.4c) produced different results when using identical random number seeds but different versions of the boost library (www.boost.org, v1.33 and v1.40) because of code changes in the random number generator implementation (supplementary section X, Supplementary Material online). We observed this behavior by pure chance while assembling a respective archive that will allow for fully reproducing our results (http://exelixis-lab.org/pavlos/gopopgen.tar.gz).

Conclusion

Literature mining for constructing narratives requires a considerable amount of time. Such an investment is useful as long as we are convinced that the results are not an artifact either of the analysis approach (e.g., the outlier approach may produce spurious results) or of model misspecification, which may result in rejecting an unrealistic null hypothesis, or of other possible sources of errors. Given that neutrality tests entail a large number of simplifying assumptions, we should focus more on analyzing method behavior rather than on interpreting or "plausibilifying" the results. Furthermore, genome scan-based selection analyses suffer from the absence of an a priori hypothesis. Thus, the number of potential interpretations of the results is practically unlimited. A well-designed experimental setup should therefore strive to reduce the number of hypotheses a priori such as to simplify the interpretation of the results. For instance, combining QTL scans with selection mapping could be deployed to narrow down the plethora of potential hypotheses to those that are associated with the trait under consideration. However, given that QTL span large genomic regions we should once again be cautious with interpretations. Alternatively, a list of candidate genes could be obtained prior to the analysis based on experimental evidence. In this case, it would be more straightforward to test if neutrality tests detect genes from the list and how often the candidate genes appear in the tail of the distribution from the null hypothesis. Therefore, we suggest that result-selling narratives in biology should be abandoned in favor of a more critical re-examination of the results.

Supplementary Material

Supplementary sections I–X are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We thank Jüri Reimand (University of Toronto) and Tambet Arak (University of Tartu) for integrating *Drosophila* annotations to g:GOSt directly from the Gene Ontology project (www.geneontology.org) and for critical comments on the manuscript. We also thank two reviewers, Vanessa Bauer DuMont and Zheng Wang (Cornell University), for their valuable suggestions. J.D.J. has been funded by the EPFL, a grant from the National Science Foundation (DEB-1002785), and a faculty award from the Worcester Foundation. W.S. has been supported by a Research Unit grant (STE 325/12) from the German Research Foundation.

References

Akey JM, Zhang G, Zhang K, Jin L, Shriver MD. 2002. Interrogating a high-density SNP map for signatures of natural selection. *Genome Res* 12:1805–1814.

Al-Shahrour F, Minguez T, Marqués-Bonet T, Gazave E, Navarro A, Dopazo J. 2010. Selection upon genome architecture: conservation of functional neighborhoods with changing genes. *PLoS Comput Biol* 6:e1000953.

Arbeitman MN, Fleming AA, Siegal ML, Null BH, Baker BS. 2004. A genomic analysis of *Drosophila* somatic sexual differentiation and its regulation. *Development* 131:2007–2021.

Ayroles JF, et al. (11 co-authors). 2009. Systems genetics of complex traits in *Drosophila melanogaster*. *Nat Genet* 41:299–307.

Barton NH. 1998. The effect of hitch-hiking on neutral genealogies. *Genetical Res* 72:123–133.

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

Beisswanger S, Stephan W, De Lorenzo D. 2006. Evidence for a selective sweep in the *wapl* region of *Drosophila melanogaster*. *Genetics* 172: 265–274.

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

Bettencourt R, Tanji T, Yagi Y, Ip YT. 2004. Toll and Toll-9 in *Drosophila* innate immune response. *J Endotoxin Res* 10:261–268.

Bustamante CD, Fiedel-Alon S, Williamson S, et al. (14 co-authors). 2005. Natural selection on protein-coding genes in the human genome. *Nature* 437:1153–1157.

Caracristi G, Schlötterer C. 2003. Genetic differentiation between American and European *Drosophila melanogaster* populations could be attributed to admixture of African alleles. *Mol Biol Evol* 20:792–799.

Carreira VP, Soto IM, Mensch J, Fanara JJ. 2011. Genetic basis of wing morphogenesis in *Drosophila*: sexual dimorphism and non-allometric effects of shape variation. *BMC Dev Biol* 11:32.

Chamilos G, Lewis RE, Hu J, Zal T, Gilliet M, Halder G, Kontoyiannis D. 2008. *Drosophila melanogaster* as a model host to dissect the immunopathogenesis of zygomycosis. *Proc Natl Acad Sci U S A* 105: 9367–9372.

Chen GK, Marjoram P, Wall JD. 2009. Fast and flexible simulation of DNA sequence data. *Genome Res* 19:136–142.

Cruickshank T, Nista P. 2011. Selection and constraint on regulatory elements in *Drosophila simulans*. *J Mol Evol* 73:94–100.

Depaulis F, Mousset S, Veuille M. 2003. Power of neutrality tests to detect bottlenecks and hitchhiking. *J Mol Evol* 57(Suppl 1), S190–200.

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

Flicek P, Amode M, Barrell D, et al. (50 co-authors). 2011. Ensembl 2011. *Nucleic Acids Res* 39:D800–806.

Galtier N, Depaulis F, Barton NH. 2000. Detecting bottlenecks and selective sweeps from DNA sequence polymorphism. *Genetics* 155: 981–987.

Gu J, Orr N, Park SD, Katz LM, Sulimova G, MacHugh DE, Hill EW. 2009. A genome scan for positive selection in thoroughbred horses. *PloS One* 4:e5767.

Haddrill PR, Thornton KR, Charlesworth B, Andolfatto P. 2005. Multilocus patterns of nucleotide variability and the demographic

and selection history of *Drosophila melanogaster* populations. *Genome Res* 15:790–799.

Harbison ST. 2004. Quantitative trait loci affecting starvation resistance in *Drosophila melanogaster*. *Genetics* 166:1807–1823.

Hoffmann R, Valencia A. 2004. A gene network for navigating the literature. *Nat Genet* 36:664.

Hoffmann R, Valencia A. 2005. Implementing the iHOP concept for navigation of biomedical literature. *Bioinformatics (Oxford, England)* 21(Suppl 2), ii252–258.

Jensen JD, Bauer DuMont VL, Ashmore AB, Gutierrez A, Aquadro CF. 2007. Patterns of sequence variability and divergence at the diminutive gene region of *Drosophila melanogaster*: complex patterns suggest an ancestral selective sweep. *Genetics* 177:1071–1085.

Jensen JD, Kim Y, DuMont VB, Aquadro CF, Bustamante CD. 2005. Distinguishing between selective sweeps and demography using DNA polymorphism data. *Genetics* 170:1401–1410.

Kayser M, Brauer S, Stoneking M. 2003. A genome scan to detect candidate regions influenced by local natural selection in human populations. *Mol Biol Evol* 20:893–900.

Kelley JL, Madeoy J, Calhoun JC, Swanson W, Akey JM. 2006. Genomic signatures of positive selection in humans and the limits of outlier approaches. *Genome Res.* 16:980–989.

Kim Y, Nielsen R. 2004. Linkage disequilibrium as a signature of selective sweeps. *Genetics* 167:1513–1524.

Kim Y, Stephan W. 2002. Detecting a local signature of genetic hitchhiking along a recombining chromosome. *Genetics* 160:765–777.

Langton PF, Colombani J, Chan EHY, Wepf A, Gstaiger M, Tapon N. 2009. The dASPP-dRASSF8 complex regulates cell-cell adhesion during *Drosophila* retinal morphogenesis. *Curr Biol* 19:1969–7198.

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

Leulier F, Lhocine N, Lemaitre B, Meier P. 2006. The *Drosophila* inhibitor of apoptosis protein DIAP2 functions in innate immunity and is essential to resist gram-negative bacterial infection. *Mol Cell Biol* 26: 7821–7831.

Li H, Stephan W. 2006. Inferring the demographic history and rate of adaptive substitution in *Drosophila*. *PLoS Genet* 2:e166.

Li J, Zhang L, Zhou H, Stoneking M, Tang K. 2011. Global patterns of genetic diversity and signals of natural selection for human *ADME* genes. *Hum Mol Genet* 20:528–540.

Li X, Carthew RW. 2005. A microRNA mediates EGF receptor signaling and promotes photoreceptor differentiation in the *Drosophila* eye. *Cell* 123:1267–1277.

McVean GAT, Cardin NJ. 2005. Approximating the coalescent with recombination. *Philos Trans R Soc London Ser B: Biol Sci* 360: 1387–1393.

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

Nielsen R, Williamson S, Kim Y, Hubisz M, Clark A, Bustamante C. 2005. Genomic scans for selective sweeps using SNP data. *Genome Res* 15: 1566–1575.

Ohh M. 2006. pVHL's kryptonite: E2-EPF UCP. *Cancer Cell* 10:95–97.

Oleksyk TK, Smith MW, O'Brien SJ. 2010. Genome-wide scans for footprints of natural selection. *Philos Trans R Soc London Ser B: Biol Sci* 365:185–205.

Pascual M, Chapuis MP, Mestres F, Balanyà J, Huey RB, Gilchrist GW, Serra L, Estoup A. 2007. Introduction history of *Drosophila subobscura* in the New World: a microsatellite-based survey using ABC methods. *Mol Ecol* 16:3069–3083.

Pavlidis P, Jensen JD, Stephan W. 2010. Searching for footprints of positive selection in whole-genome SNP data from non-equilibrium populations. *Genetics* 185:907–922.

Pool JE, Hellmann I, Jensen JD, Nielsen R. 2010. Population genetic inference from genomic sequence variation. *Genome Res* 20:291–300.

Przeworski M. 2002. The signature of positive selection at randomly chosen loci. *Genetics* 160:1179–1189.

Reimand J, Arak T, Vilo J. 2011. g:Profiler—a web server for functional interpretation of gene lists (2011 update). *Nucleic Acids Res* 39: W307–315.

Reimand J, Kull M, Peterson H, Hansen J, Vilo J. 2007. g:Profiler–a web-based toolset for functional profiling of gene lists from large-scale experiments. *Nucleic Acids Res* 35:W193–200.

Rubin C-J, Zody M, Eriksson J, et al. (19 co-authors). 2010. Whole-genome resequencing reveals loci under selection during chicken domestication. *Nature* 464:587–591.

Sambandan D, Carbone MA, Anholt RRH, Mackay TFC. 2008. Phenotypic plasticity and genotype by environment interaction for olfactory behavior in *Drosophila melanogaster*. *Genetics* 179: 1079–1088.

Sella G, Petrov DA, Przeworski M, Andolfatto P. 2009. Pervasive natural selection in the *Drosophila* genome? *PLoS Genetics* 5:e1000495.

Stamatakis A, Izquierdo-Carrasco F. 2011. Result verification, code verification and computation of support values in phylogenetics. *Brief Bioinform* 12:270–279.

Stark C, Breitkreutz B, Chatr-Aryamontri A, et al. (15 co-authors). 2011. The BioGRID Interaction Database: 2011 update. *Nucleic Acids Res* 39:D698–704.

Stephan W. 2010a. Detecting strong positive selection in the genome. *Mol Ecol Res* 10:863–872.

Stephan W. 2010b. Genetic hitchhiking versus background selection: the controversy and its implications. *Philos Trans R Soc London Series B: Biol Sci* 365:1245–1253.

Svetec N, Pavlidis P, Stephan W. 2009. Recent strong positive selection on *Drosophila melanogaster HDAC6*, a gene encoding a stress surveillance factor, as revealed by population genomic analysis. *Mol Biol Evol* 26:1549–1556.

Svetec N, Werzner A, Wilches R, Pavlidis P, Alvarez-Castro J, Broman KW, Metzler D, Stephan W. 2011. Identification of X-linked quantitative trait loci affecting cold tolerance in *Drosophila melanogaster* and fine mapping by selective sweep analysis. *Mol Ecol* 20: 530–544.

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

Tauber E, Zordan M, Sandrelli F, et al. (13 co-authors). 2007. Natural selection favors a newly derived timeless allele in *Drosophila melanogaster*. *Science* 316:1895–1898.

Thornton K, Andolfatto P. 2006. Approximate Bayesian inference reveals evidence for a recent, severe bottleneck in a Netherlands population of *Drosophila melanogaster*. *Genetics* 172:1607–1619.

Thornton KR, Jensen JD. 2007. Controlling the false-positive rate in multilocus genome scans for selection. *Genetics* 175:737–750.

Torgerson DG, Boyko A, Hernandez R, et al. (11 co-authors). 2009. Evolutionary processes acting on candidate cis-regulatory regions in humans inferred from patterns of polymorphism and divergence. *PLoS Genet* 5:e1000592.

Vastrik I, D'Eustachio P, Schmidt E, et al. (14 co-authors). 2007. Reactome: a knowledge base of biologic pathways and processes. *Genome Biol* 8:R39.

Voight BF, Kudaravalli S, Wen X, Pritchard JK. 2006. A map of recent positive selection in the human genome. *PLoS Biol* 4:e72.

Wijnen H, Naef F, Boothroyd C, Claridge-Chang A, Young MW. 2006. Control of daily transcript oscillations in Drosophila by light and the circadian clock. *PLoS Genet* 2:e39.

Williamson SH, Hubisz M, Clark A, Payseur BA, Bustamante C, Nielsen R. 2007. Localizing recent adaptive evolution in the human genome. *PLoS Genet* 3:e90.

Wu T, Merbl Y, Huo Y, Gallop JL, Tzur A, Kirschner MW. 2010. UBE2S drives elongation of K11-linked ubiquitin chains by the anaphase-promoting complex. *Proc Natl Acad Sci U S A* 107: 1355–1360.

Young MD, Wakefield MJ, Smyth GK, Oshlack A. 2010. Gene ontology analysis for RNA-seq: accounting for selection bias. *Genome Biol* 11: R14; doi:10.1186/gb-2010-11-2-r14.

Yukilevich R, Turner TL, Aoki F, Nuzhdin SV, True JR. 2010. Patterns and processes of genome-wide divergence between North American and African *Drosophila melanogaster*. *Genetics* 186: 219–239.