DOI: 10.1534/genetics.110.123059

Genomic Differentiation Between Temperate and Tropical Australian Populations of *Drosophila melanogaster*

Bryan Kolaczkowski,*,[1] Andrew D. Kern,*,[1] Alisha K. Holloway† and David J. Begun†,[2]

**Department of Biological Sciences, Dartmouth College, Hanover, New Hampshire 03755 and †Department of Evolution and Ecology, University of California, Davis, California 95616*

Manuscript received September 9, 2010
Accepted for publication November 3, 2010

ABSTRACT

Determining the genetic basis of environmental adaptation is a central problem of evolutionary biology. This issue has been fruitfully addressed by examining genetic differentiation between populations that are recently separated and/or experience high rates of gene flow. A good example of this approach is the decades-long investigation of selection acting along latitudinal clines in *Drosophila melanogaster*. Here we use next-generation genome sequencing to reexamine the well-studied Australian *D. melanogaster* cline. We find evidence for extensive differentiation between temperate and tropical populations, with regulatory regions and unannotated regions showing particularly high levels of differentiation. Although the physical genomic scale of geographic differentiation is small—on the order of gene sized—we observed several larger highly differentiated regions. The region spanned by the cosmopolitan inversion polymorphism *In(3R)P* shows higher levels of differentiation, consistent with the major difference in allele frequencies of Standard and *In(3R)P* karyotypes in temperate *vs.* tropical Australian populations. Our analysis reveals evidence for spatially varying selection on a number of key biological processes, suggesting fundamental biological differences between flies from these two geographic regions.

DETERMINING the processes maintaining genetic variation within species is a basic goal of biological research and a central problem of evolutionary genetics. Indeed, the relative contributions to segregating variation of (1) low-frequency, unconditionally deleterious mutations, (2) intermediate-frequency, small-effect variants maintained by mutation and genetic drift, and (3) adaptive mutations maintained by positive selection—*e.g.*, spatially varying or negative frequency-dependent selection—remain unknown in any species. Thus, it is also unclear whether different processes predominate in different species, perhaps resulting from differences in population size, ecology, or genetics.

One approach for identifying adaptive variants segregating within species is to investigate systems in which there are major phenotypic variants likely influenced by natural selection and that have relatively simple genetics. This is what has traditionally been thought of as ecological genetics. For example, pigmentation variation in vertebrates (*e.g.*, Nachman *et al.* 2003) is a good example of a trait for which the relatively small number of candidate genes allows the phenotypic effects of natural variants to be directly tested. For major phenotypic variants having a simple genetic basis but no candidate genes, genetic analysis can be used to isolate alternative alleles underlying the phenotypic difference. Examples include diapause variation and foraging behavior in *Drosophila melanogaster* (Osborne *et al.* 1997; Schmidt *et al.* 2008), traits relating to social behavior and copulatory plug formation in *Caenorhabditis elegans* (de Bono and Bargmann 1998; Palopoli *et al.* 2008), and several phenotypes in sticklebacks (Colosimo *et al.* 2004; Miller *et al.* 2007; Chan *et al.* 2010). Besides their simple genetics, such biological examples have the advantage that the targeted traits may have plausible connections to fitness variation in nature (though this is not always the case). In spite of the practical advantages associated with phenotypic variation resulting from simple genetics and alleles of large effect, such variation may not speak very strongly to the general properties of adaptive polymorphisms in natural populations, which may often be characterized by complex genetics or small-effect alleles.

A complementary approach uses population-genetic analysis to identify individual polymorphic variants/genes that may have been influenced by positive selection. This approach offers at least two advantages. First, it can be made genomic in scope and therefore may provide a less-biased view of the genes and phenotypes influenced by positive selection. There is no comparably comprehensive "omic" concept for phenotypic analysis, because

Supporting information is available online at http://www.genetics.org/cgi/content/full/genetics.110.123059/DC1.

[1]These authors contributed equally to this work.

[2]*Corresponding author:* Department of Evolution and Ecology, College of Biological Sciences, University of California, 3350A Storer, Davis, CA 95616. E-mail: djbegun@ucdavis.edu

the universe of phenotype space is difficult to define, difficult to measure, and highly dimensional (Lewontin 1974). Second, alleles having relatively small effects or effects not associated with easily defined phenotypes can be identified. A population-genetic approach is a particularly powerful discovery tool when joined with high-quality genome annotation, generating many new hypotheses about the genetic and phenotypic variation influenced by positive selection within species and providing vast opportunities for the downstream functional investigation of such variation.

One population-genetic approach for identifying positively selected polymorphisms is to search the genome for sites exhibiting large allele-frequency differences between recently separated populations or those experiencing high rates of gene flow (Lewontin and Krakauer 1975). Because even low levels of gene flow effectively homogenize neutral allele frequencies (Wright 1931; Maruyama 1970; Slatkin 1981), alleles under spatially varying selection are expected to appear as outliers with respect to allele-frequency differences across populations. This strategy may be particularly effective when allele frequencies change gradually along a cline, such as with latitude or altitude.

Some of the best-studied cases of latitudinal clines maintained by spatially varying selection are those of *D. melanogaster*. The majority of work on these clines has investigated various phenotypic traits, chromosome inversion polymorphisms, and enzyme-coding genes (Sezgin *et al.* 2004), as well as several other genes harboring clinal variants (Costa *et al.* 1992; McColl and McKechnie 1999; Schmidt *et al.* 2000; Duvernell *et al.* 2003). The cline along the east coast of Australia has received considerable recent attention due to the efforts of Ary Hoffmann and collaborators (*e.g.*, Hoffmann and Weeks 2007). The fact that similar clines are often observed on different continents strongly implicates natural selection rather than demography as the cause of clinal variation (Oakeshott *et al.* 1981, 1983; Singh and Rhomberg 1987; Singh 1989; Singh and Long 1992; Gockel *et al.* 2001; Kennington *et al.* 2003; Hoffmann and Weeks 2007). Importantly, although cosmopolitan chromosome inversion polymorphisms exhibit latitudinal clines (with inversion frequency increasing in more tropical populations), many observations convincingly show that inversions explain only a fraction of clinal variation, even for genes located in inverted regions (Voelker *et al.* 1978; Knibb 1982; Singh and Rhomberg 1987; Frydenberg *et al.* 2003; Umina *et al.* 2006). Indeed, many clinally varying genes are not physically near inversions (Voelker *et al.* 1978; Singh and Rhomberg 1987; Sezgin *et al.* 2004; Turner *et al.* 2008).

We recently extended the genetic characterization of population differentiation from *D. melanogaster* clines by comparative genomic hybridization analysis of population samples from opposite ends of well-described clines in Australia and North America (Turner *et al.* 2008). That study generated new information on genomic differentiation, but the crude nature of the data limited the scope of the analysis and the strength of the conclusions that could be drawn. Here we revisit the issue of geographic differentiation between opposite ends of a known *D. melanogaster* cline, using next-generation sequencing to characterize genomic variation in flies from Queensland and Tasmania, Australia. These data are used to generate hypotheses regarding the biological differences between flies from these regions and to assess the population-genetic properties of sequence differentiation between these geographic regions.

MATERIALS AND METHODS

Sequencing, assembly, and data filtering: Population samples from the east coast of Australia were collected in 2004 (Anderson *et al.* 2005). Twenty isofemale lines from Queensland (Cairns, lat. 16.907, and Cooktown, lat. 15.476) and 19 isofemale lines from Tasmania (Hillwood, lat. 41.237, and Sorell, lat. 42.769) were used. Two females were collected from each Queensland line ($n = 40$ flies). These flies were pooled in a single tube and made into DNA. Similarly, two females were collected from each Tasmania line ($n = 38$ flies), pooled in a single tube, and made into DNA. Each of the two DNA samples was then sequenced using Solexa/Illumina technology (Bentley *et al.* 2008). Base calls and quality scores were determined using the Solexa GAPipeline v0.3.0. Output files were in fastq format. Reads were mapped against the *D. melanogaster* reference genome R5.8 (Adams *et al.* 2000), using Maq v0.6.8 (Li *et al.* 2008). Prior to mapping, we split fastq files into separate files with 1 million reads per file. The reads are available in the NCBI Sequence Read Archive under accession no. SRA012285.16.

Several Maq functions were used for data formatting. Solexa quality scores were converted to Sanger quality scores using Maq function sol2sanger and converted from fastq files to binary fastq (bfq) using the Maq function fastq2bfq. Bases 1–36 of each read were used; the expected heterozygosity parameter ("$-m$" flag) was 0.005. Mapped reads were merged using mapmerge. The functions maq assemble and maq pileup were then used to produce pileup files. Finally, pileup files were split by chromosome arm for downstream analysis. Individual base calls with Maq quality scores <10 were excluded, as were positions with only a singleton variant in the entire Australian sample. We explored the value of increasing the Maq quality threshold to 20, but the reduction in coverage was too costly, given the amount of data. Because we excluded singletons and focused on genomic outliers, errors should not be an important factor with respect to our biological conclusions. We excluded genomic positions with <6 or >20 sequence reads in either population, because these sites are associated either with very low power to reject the null hypothesis or with the confounding phenomenon of differentiated copy-number variation.

Because a primary goal of our study was to generate biological, gene-centric hypotheses regarding the nature of selection, most analyses excluded regions of the genome adjacent to centromeres and telomeres associated with low heterozygosity, as determined from genome sequences of a Raleigh sample of inbred lines sequenced as part of the Drosophila Population Genomics Project (DPGP.org). These regions of reduced heterozygosity are expected to be associated with lower power to detect differentiation, and because

they experience reduced rates of crossing over, the physical scale of differentiation may be quite large, limiting opportunities for identifying potential targets of selection. The coordinates corresponding to regions of normal recombination used in our analyses are *2*L, 844,225–19,946,732; *2*R, 6,063,980–20,322,335; *3*L, 447,386–18,392,988; *3*R, 7,940,899–27,237,549; and *X*, 1,036,552–20,902,578. The regions excluded are roughly consistent with the non- or low-recombining portions of the genome identified in prior studies (*e.g.*, Singh *et al.* 2005).

Ancestral sequence reconstruction: For the purposes of unfolding the site frequency spectrum in our samples, ancestral states were inferred using maximum likelihood (ML) (Yang *et al.* 1995) [provided by PAML v4.3 (Yang 2007)], assuming the reference phylogeny (Drosophila 12 Genomes Consortium 2007), the HKY nucleotide substitution model (Hasegawa *et al.* 1985), and gamma-distributed among-site rate variation (Yang 1996). ML reconstruction posterior probabilities were calculated using the empirical Bayesian approach described in Yang *et al.* (1995); the posterior probability of ancestral base b_i, given data x_j at alignment position j, is given by $P(b_i|x_j) = P(x_j|b_i)P(b_i)/\sum_{k=1}^{4} P(x_j|b_k)P(b_k)$, where $P(x_j|b_i)$ is the probability of observing data x_j given base b_i in the ancestral sequence, and $P(b_i)$ is the frequency of base b_i in the data set. Positions with a ML reconstruction posterior probability <0.9 were considered potentially unreliable and excluded from the analysis. The data for our ancestral sequence reconstruction were obtained from the MULTIZ 15-way insect alignment available for download from the UCSC genome browser (Blanchette *et al.* 2004; Hinrichs *et al.* 2006).

Population genetic estimation of pooled sample reads: Although the pooling strategy provides an economical picture of sequence polymorphism, it is associated with atypical sampling properties. Here we provide results for bias-corrected estimators of heterozygosity and other canonical population genetic summary statistics.

Sequencing pooled DNA leads to an additional round of sampling with replacement, beyond the initial sampling of chromosomes from nature. Let p be the population frequency of an allele A_1. Also consider the case where n chromosomes are sampled from nature and are sequenced to a depth m. We do not treat m as a random variable, although other authors have (Futschik and Schlotterer 2010). The probability of sequencing $X = k$ from m reads of the A_1 allele, conditional upon the population frequency p and our pooled sample size n, is

$$\begin{aligned}\text{Prob}(X = k \mid m, n, p) &= \sum_{i=0}^{n} \binom{m}{k}\left(\frac{i}{n}\right)^k\left(1-\frac{i}{n}\right)^{m-k}\binom{n}{i}p^i(1-p)^{n-i}. \qquad (1)\end{aligned}$$

The expected value of the sample frequency, $E(k/m)$, should be unbiased with respect to the frequency in the population, as $E(k/m) = E(E(k/m|i/n)) = \sum_i E(k/m|i/n) \times \text{Prob}(i) = p$. Deriving the second moment of the sample frequency is more involved and can be found in supporting information, File S1. The result is $E((k/m)^2) = p(1-p)(n-1+m)/nm + p^2$, which allows one to write down an unbiased estimator of heterozygosity $H = 2p(1-p)$. Under standard binomial sampling, the estimator H is biased and needs to be corrected by a factor of $n/(n-1)$ (Nei 1987). In the case of sequencing into pooled samples, the expectation of H is

$$\begin{aligned}E(H) = E(2p(1-p)) &= 2(E(p) - E(p^2)) \\ &= 2p(1-p)\left(\frac{n-1}{n}\right)\left(\frac{m-1}{m}\right). \qquad (2)\end{aligned}$$

The correction for the second round of sampling adds one term to the estimator of heterozygosity. The correction leads to our estimate of allele-frequency differentiation between Queensland and Tasmania, F_{ST}, which was calculated as

$$F_{ST} = \frac{\Pi_{total} - \Pi_{within}}{\Pi_{total}},$$

where

$$\begin{aligned}\Pi_{total} &= H(P_{total}) \\ \Pi_{within} &= \frac{(N_Q \times H(P_Q)) + (N_{TAS} \times H(P_{TAS}))}{N_Q + N_{TAS}} \\ H(P) &= 2p(1-p)\frac{n}{n-1}\frac{m}{m-1}.\end{aligned}$$

Here N_Q and N_{TAS} are the sample sizes from Queensland and Tasmania populations, respectively, and P_Q and P_{TAS} are the corresponding allele frequencies. P_{total} is the allele frequency in the combined (*i.e.*, Queensland and Tasmania) population sample. $H(P)$ is our corrected estimate of heterozygosity from Equation 2. In File S1 we provide simulation results that demonstrate our corrected version of F_{st} is unbiased with respect to coverage.

Estimators of θ: As above in our treatment of heterozygosity, we need to correct estimators of the neutral mutation parameter $\theta = 4Nu$ for a pooled sampling strategy. Some recent work on this problem was done by Futschik and Schlotterer (2010), who consider the case of pooled samples when the pool is large in comparison to sequence coverage. Here and in File S1, we derive results for corrected estimators that are accurate in the case where coverage is of similar size to the pooled sample. Importantly, we can derive the expected site frequency spectrum of a pooled sequencing experiment.

The first result of interest is the probability of observing an allele segregating at frequency k from m in our sequenced sample, given a pooled sample size of n. This will differ from the quantity in Equation 1, because we sum over possible allele frequencies of the A_1 allele in the sample, i, in accordance with their expected probabilities under the standard neutral model. Thus the unconditional probability is

$$\begin{aligned}\text{Prob}(k \mid m, n) &= \sum_{i=1}^{n-1} \text{Prob}(k \mid m, n, i)\text{Prob}(i) \\ &= \sum_{i=1}^{n-1}\binom{m}{k}\left(\frac{i}{n}\right)^k\left(1-\frac{i}{n}\right)^{m-k}\left(\frac{1}{ia_n}\right), \qquad (3)\end{aligned}$$

where $a_n = \sum_{j=1}^{n-1} 1/j$. The last term in Equation 3 represents the probability of observing an allele segregating at frequency i from n chromosomes under the neutral model (Ewens 2004). Fu (1995) was able to derive the expected number of sites, X_i segregating at frequency i from n as $E\{X_i\} = \theta/i$. While Fu derived his result from modeling the genealogical process as a form of the Polya urn model, a simpler derivation comes by conditioning on the total number of segregating sites in a sample, S. Conditional on S, the X_i's can be assumed to follow a multinomial distribution where the individual parameters reflect the expected frequencies of sites in the sample. Using this logic, $E\{X_i\} = E\{S\} \times \text{prob}(i) = \theta a_n \times 1/ia_n = \theta/i$. Similarly we can write the expected counts of each frequency class in our sequenced sample Y_i,

$$\begin{aligned}E\{Y_k\} &= E\{S\} \times \text{Prob}(k \mid m, n) \\ &= \theta a_n \sum_{i=1}^{n-1}\binom{m}{k}\left(\frac{i}{n}\right)^k\left(1-\frac{i}{n}\right)^{m-k}\left(\frac{1}{ia_n}\right). \qquad (4)\end{aligned}$$

We point the reader to File S1 for simulation results confirming the accuracy of this expression. With the expected site

frequency spectrum in hand, we can use the weighted linear combination of Achaz (2009) to write down estimators of θ given our sampling regime. In particular, given the high sequencing error rates inherent in these data, we want modified estimators of θ that exclude singletons.

Modified versions of Tajima's nucleotide diversity $\hat{\theta}_\pi$ and Fay and Wu's $\hat{\theta}_H$ (Tajima 1983; Fay and Wu 2000) were computed as follows. Let Y_k represent the number of sites segregating in a region at derived frequency k from m reads, given a pool of n chromosomes. One can write an unbiased estimator of θ using arbitrary weights for each frequency class ω_i, such that

$$\hat{\theta}_\omega = \frac{1}{a_n \sum_k \omega_k} \sum_{k=1}^{m-1} \omega_k Y_k \frac{1}{\text{Prob}(k \mid m, n)}. \quad (5)$$

This result allows for generalized weighted estimators of θ given pooled sampling. We present simulation results in File S1 that demonstrate our new estimators are accurate and unbiased with respect to coverage. In the present case, we are interested in two weighting schemes, one to create a modified $\hat{\theta}_\pi$ and the other for a modified $\hat{\theta}_H$ estimator. Let the associated weights be $\omega_{\pi,k}$ and $\omega_{H,k}$, respectively. Then

$$\omega_{\pi,k} = \begin{cases} 0 & k = 1 \\ m - k & 1 < k \leq m - 1 \end{cases}$$

and

$$\omega_{H,k} = \begin{cases} 0 & k = 1 \\ k & 1 < k \leq m - 1. \end{cases}$$

The modified Fay and Wu's H that excludes singleton sites is the difference between our two estimators. As our estimators are unbiased with respect to coverage, $\hat{\theta}$ over a region where m (coverage) varies is simply the sum of $\hat{\theta}$ at each m.

Outlier approach: The relative merit of a model-based inference from theory or simulations *vs.* an empirical genomic-based outlier approach for detecting targets of positive selection is an ongoing discussion in the literature (Beaumont and Nichols 1996; Akey *et al.* 2002; Beaumont and Balding 2004; Teshima *et al.* 2006; Voight *et al.* 2006; Pickrell *et al.* 2009). For the following reasons, we chose to use an empirically based outlier approach for identifying candidate targets of selection: (1) the challenges associated with generating a realistic null model for our *D. melanogaster* cline are substantial, (2) we have relatively few data from which to estimate model parameters, (3) there is little doubt that many of the highly differentiated genomic regions from the east Australian cline result from selection, and (4) the empirical approach represents a simple, transparent treatment of the data. The many consistent biological signals we report here support the value of this approach, although they do not speak to its optimality.

Because the true length distribution of differentiated regions is unknown, two main approaches were used to identify such regions. Mean F_{ST} values were calculated for 1-kb nonoverlapping windows across the normally recombining regions of the genome. The top 1% or top 2.5% of these windows were considered "differentiated" for most analyses. For some analyses, the 5% tail was used (see Figure S1a and results section below). To identify differentiation on a scale >1 kb, we aggregated 1-kb windows in our top 1% tail. We considered any region of at least five consecutive windows that were not in the top 10% of mean 1-kb F_{ST} as "undifferentiated" between Queensland and Tasmania. Any region between two undifferentiated regions that had at least one 1-kb window in the top 1% F_{ST} was considered an independent differentiated region. We additionally investigated very small-scale differentiation by considering the top 0.1% of individual-position F_{ST} values not occurring in the top 10% 1-kb windows as potential outlier variants. Unless otherwise noted, all analyses were restricted to outliers occurring in normally recombining regions.

Genome annotations were taken from FlyBase R5.24 (Tweedie *et al.* 2009). Genome positions were annotated as coding sequence (CDS), 3′- and 5′-UTR, intron, regulatory, and "other." Because regulatory regions are underrepresented in the FlyBase annotation, additional regulatory annotations were retrieved from the OregAnno database (Griffith *et al.* 2008) and a recent genome-wide scan for transcription-factor binding sites (MacArthur *et al.* 2009). Polymorphisms within coding sequence were additionally annotated as either nonsynonymous or synonymous.

Gene Ontology (GO) annotations (Ashburner *et al.* 2000) were obtained from FlyBase R5.24 (Tweedie *et al.* 2009). For each GO annotation, the number of genes within all 1-kb normally recombining windows with that annotation were identified. GO-category enrichment was determined using a hypergeometric test that compared the proportion of genes with a given GO annotation to the proportion of genes in the 2.5% most-differentiated 1-kb windows with that GO annotation. All GO categories with fewer than four genes were excluded, as four genes are the minimum number for which a significant hypergeometric result is possible at $\alpha = 0.05$. After controlling the false discovery rate using the method of Storey (2002), enriched GO categories with false discovery rate (FDR)-corrected P-values <0.05 were determined. Similar GO-category enrichment analyses were performed using individual outlier genomic positions. Of course, differentiation at specific genes could have profound phenotypic consequences without leaving a statistically significant signature of GO enrichment.

Copy-number variation was evaluated by calculating the mean coverage for nonoverlapping 1-kb windows across Queensland and Tasmania genomes. For each window, we calculated the ratio of Queensland/Tasmania coverage and normalized these ratios by the mean coverage ratio across each chromosome arm. The top 1, 2.5, and 5% most-extreme windows were considered highly differentiated in copy number (see Figure S1b). Gene Ontology enrichment analyses were conducted as described above.

Structure prediction: RNA secondary structures were inferred using the Vienna RNA package v1.8.2 (Hofacker 2003) with default parameters. Protein domain architecture was inferred using a sequence search of the PFam database (Coggill *et al.* 2008; Finn *et al.* 2010). Homology-based 3D structural modeling was performed using MODELER 9v7 (Eswar *et al.* 2008). Structures were inferred for predicted proteins from a consensus sequence for Queensland and Tasmania genes *Irc* and *NtR*. Searching the Protein Data Bank (Berman *et al.* 2000) using *melanogaster* protein sequences returned structures 3ERH (Sheikh *et al.* 2009) and 2QC1 (Dellisanti *et al.* 2007) as the best matches to the predicted proteins of *Irc* and *NtR*, respectively. Queensland and Tasmania consensus protein sequences were aligned to each structural template using MAFFT v6.611 with the E-INS-i option (Katoh *et al.* 2002; Katoh and Toh 2008). Five structural models of each sequence were constructed and evaluated using the MODELER objective function as well as DOPE and GA341 assessment scores (Eramian *et al.* 2008). Results are shown for the best overall models. Sequence not alignable to the structural template was excluded.

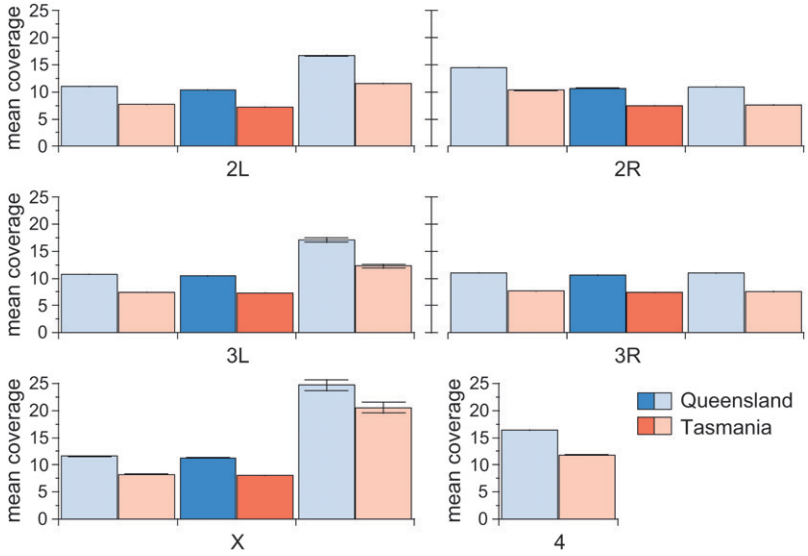

FIGURE 1.—Genome-sequence coverage is equivalent across chromosome arms in normally recombining regions and more variable in low-recombining regions. Mean sequencing coverage is plotted for Queensland (blue) and Tasmania (red) populations. Dark colors indicate regions of normal recombination; lighter colors indicate low-recombining centromeric and telomeric regions. Bars give standard error.

RESULTS

After filtering, the average genome coverage was 11.6× in Queensland and 8.2× in Tasmania. Coverage varied little across chromosome arms (Figure 1). The Queensland/Tasmania coverage ratio was highly consistent, varying from 1.20 to 1.45 across all regions examined. In addition, coverage in normally recombining regions was nearly equivalent across chromosome arms: the *X* chromosome had the greatest coverage (11.3 and 8.0 in Queensland and Tasmania, respectively), while chromosome *2*L had the lowest (10.4 and 7.3). After filtering, the mean coverage and mean number of SNPs per 1-kb window were 604.7 bp and 9.4, respectively.

Genomic patterns: Mean F_{ST} across the entire genome was $0.112 \pm 8.23 \times 10^{-5}$. The distribution of 1-kb window F_{ST} estimates has a long right tail (see Figure S1a); the 5, 2.5, and 1% thresholds for this tail are $F_{ST} = 0.23$, $F_{ST} = 0.27$, and $F_{ST} = 0.32$, respectively. Among-arm variation in F_{ST} was significantly heterogeneous (Kruskal–Wallis rank sum test: $P < 2.2 \times 10^{-16}$; see also Table S1); the rank order of mean F_{ST} across chromosome arms was *3R* (0.124) > *2L* (0.116) > *3L* (0.111) > *2R* (0.107) > *X* (0.097). Previous studies demonstrated that *In(3R)P* *vs.* Standard represents a nearly fixed difference between Queensland and Tasmania (corresponding to F_{ST} close to 1.0), which is considerably greater differentiation than that observed for other cosmopolitan inversions in these populations (KNIBB *et al.* 1981). This suggests that the *In(3R)P* cline is a main cause of the elevated F_{ST} for *3R*. Two aspects of the data support this proposition. First, the region spanned by *In(3R)P* was significantly more differentiated than the rest of *3R* (0.129 *vs.* 0.113, Wilcoxon's rank sum test: $P < 2.2 \times 10^{-16}$; see Figure 2c and Figure S2). Second, the physical scale of differentiation was significantly greater on chromosome arm *3R*, which exhibited slightly fewer very small differentiated regions (<2 kb) and significantly more large regions of high F_{ST} (>10 kb) compared to the other arms (Fisher's exact test, $P = 0.000378$, Figure 2b). Note that F_{ST} of nucleotide variation in the region spanned by *In(3R)P* was dramatically lower than estimates of F_{ST} of the inversion itself, based on previous studies of these populations (KNIBB *et al.* 1981; KNIBB 1982; UMINA *et al.* 2005), suggesting extensive recombination in the history of this arrangement.

In(2L)t also shows clinal variation, though not as steep as that of *In(3R)P* (KNIBB *et al.* 1981). There was also a significant difference in F_{ST} for the region spanned by *In(2L)t* (0.116) *vs.* the rest of the arm (0.109) (Wilcoxon's rank sum test: $P < 2.2 \times 10^{-16}$); however, it appears that most of the difference is explained by the region of low differentiation in the uninverted region adjacent to the centromere (see Figure S2). The other two autosomal arms similarly showed only very slightly higher F_{ST} (*3L*) or no difference in F_{ST} (*2R*) for regions spanned by cosmopolitan inversions (there is no such inversion on the *X* chromosome). Much of the difference between standard and inverted regions for arms other than *3R* is explained by reduced heterozygosity and differentiation in centromere-proximal regions that are not included in the inversions (see Figure S2).

Despite the filtering of regions corresponding to reduced heterozygosity as defined by DPGP, we observed that regions near centromeres (and some telomeres) showed low levels of differentiation, which corresponds to regions of reduced heterozygosity (see Figure S2). This suggests that some centromere- and telomere-proximal euchromatic sequence experiencing reduced crossing over may remain in our filtered data. However, the physical scale of differentiated regions was similar in normally *vs.* low-recombining regions of the genome (Figure 2a).

We detected significant heterogeneity in levels of nucleotide diversity $(\hat{\theta}_\pi)$ among chromosome arms

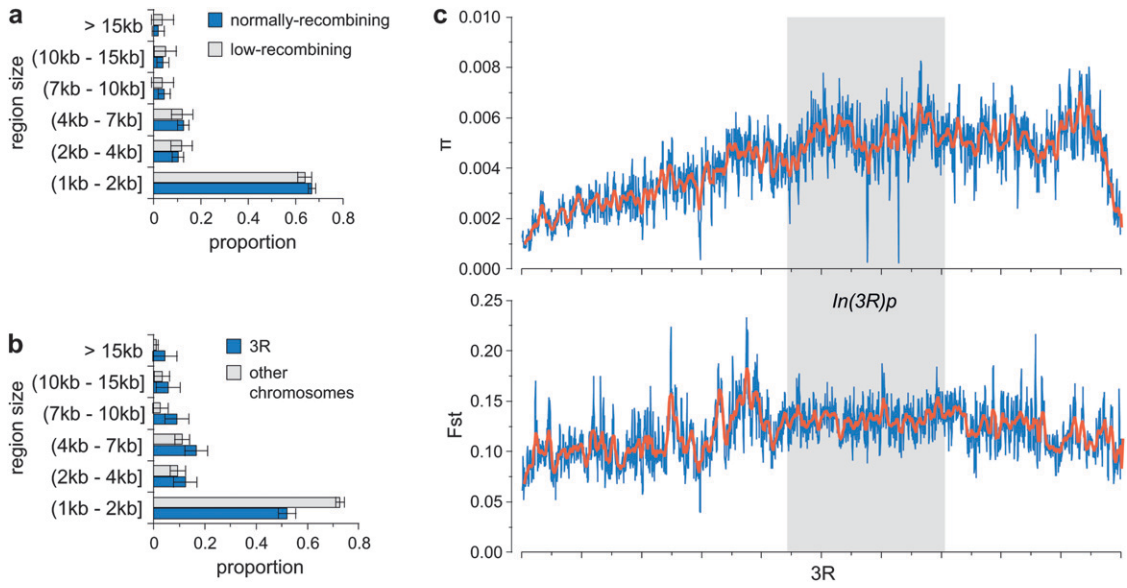

FIGURE 2.—Size of differentiated regions is similar in areas of normal and low recombination and larger on chromosome *3*R. We calculated mean F_{ST} in nonoverlapping 1-kb windows across the *D. melanogaster* genome. Groups of windows in the top 1% tail of the F_{ST} distribution were grouped together into larger differentiated regions separated from one another by at least five consecutive windows with mean F_{ST} in the bottom 90% tail (see MATERIALS AND METHODS). (a) We plot the size distribution of these differentiated regions for normally recombining (blue) and low-recombining (gray) areas of the genome. Bars indicate standard error. (b) We plot the size distribution of differentiated regions found in normally recombining regions of chromosome *3*R (blue) and the size distribution of differentiated regions in normally recombining regions of other chromosome arms (gray). (c) We plot mean F_{ST} (bottom) and mean polymorphism (π, top) across chromosome *3*R. Blue lines indicate average values over 25-kb windows slid every 10 kb; red lines show 200-kb windows slid 50 kb at a time. The gray box indicates the location of the cosmopolitan *3R-Payne* inversion.

(Kruskal–Wallis rank sum test: $P < 2.2 \times 10^{-16}$; see also Table S1), with the *X* chromosome showing the lowest diversity. We also detected systematic differences in nucleotide diversity between population samples, with the Tasmanian population showing consistently lower heterozygosity than the Queensland sample (see Table S1). Additionally, Fay and Wu's *H* statistic was significantly more negative for Tasmania than for Queensland both in the genome as a whole (Wilcoxon's rank sum test: $P < 2.2 \times 10^{-16}$; see Figure S3) and in the normally recombining portion of the genome (Wilcoxon's rank sum test: $P < 2.2 \times 10^{-16}$). One explanation for the more negative Fay and Wu's *H* statistic in Tasmania is recent strong selection in this temperate population (FAY and WU 2000). Consistent with this explanation, we found that the 1-kb regions that were very highly differentiated also exhibited considerably more negative values of *H* in Tasmania compared to Queensland, relative to the rest of the genome (Wilcoxon's rank sum tests: 5% tail, $P < 2.2 \times 10^{-16}$; 2.5% tail, $P < 2.2 \times 10^{-16}$; 1% tail, $P < 2.2 \times 10^{-16}$).

The largest differentiated euchromatic region spanned 854 kb at the tip of the *X* chromosome (Figure 3a), a region of low heterozygosity documented in several studies (AGUADE *et al.* 1989; BEGUN and AQUADRO 1995; LANGLEY *et al.* 2000). Interestingly, previous studies suggested that the scale of linkage disequilibrium in this region of the genome is not dramatically reduced, in spite of reduced levels of crossing over (BEGUN and AQUADRO 1995; LANGLEY *et al.* 2000). This suggests that differentiation at the tip of the *X* region corresponds to a mosaic linkage-disequilibrium structure of relatively low small-scale linkage disequilibrium interspersed with scattered large-scale linkage disequilibrium. The largest differentiated segment in the middle of a chromosome arm was a 752-kb region of chromosome *2*R (Figure 3b). Interestingly, *Cyp6g1*, an insecticide resistance gene (DABORN *et al.* 2002; SCHMIDT *et al.* 2010) known to be under recent strong selection, is located in this region and is an excellent candidate for the observed differentiation. Other areas of extended differentiation were observed in the euchromatic portion of the *X* chromosome (a 245-kb region from 18,055 to 18,300 kb) and toward the proximal end of chromosome *2L* (a 131-kb region from 20,172 to 20,303 kb).

The majority of differentiation between the Queensland and Tasmania populations occurs on a small physical scale (see Figure 2, a and b, and Table S1). In fact, F_{ST}-outlier regions (see MATERIALS AND METHODS) were defined by single 1-kb windows in most cases, and most such windows localize to single genes. This small-scale differentiation facilitates effective identification of candidate genes influenced by spatially varying selection. Figure 4 shows one example in which a 1-kb windows in the top 2.5% F_{ST} tail localizes to *Sfmbt*, a chromatin-binding protein involved in gene regulation (GRIMM *et al.* 2009).

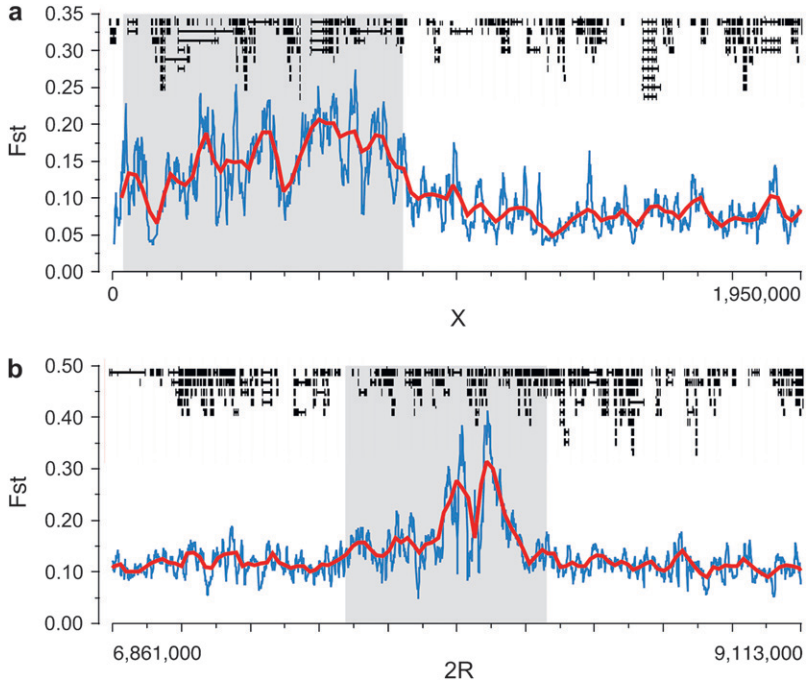

FIGURE 3.—Largest highly differentiated regions occurred at the tip of the *X* chromosome (a) and in the middle of chromosome *2R* (b). Highly differentiated regions are indicated in gray. We plot mean F_{ST} across each chromosomal region, blue lines indicating 10-kb windows with 1-kb slides and red lines indicating 50-kb windows with 20-kb slides. Annotated genes are drawn across the top of each panel.

Differentiation in this gene is primarily attributable to two fixed substitutions in the middle of the gene. Interestingly, *Sfmbt* has been shown through yeast two-hybrid studies to physically interact with seven other genes (YU *et al.* 2008), two of which—*CG33275* and *CG17018*—are also highly differentiated between Queensland and Tasmania (1-kb F_{ST} = 0.26 and 0.45, respectively). Two additional genes predicted to interact with *Sfmbt* on the basis of known interactions between human homologs—*Hdac3* and *Stam*—are also highly differentiated (1-kb F_{ST} = 0.28 and 0.33, respectively).

A genome browser displaying 1-kb windows and their associated F_{ST} estimates is available at http://altair.dartmouth.edu/ucsc/index.html. Significantly differentiated regions showed substantial overlap with outlier regions previously identified in similar Australian samples, using comparative genomic hybridization (TURNER *et al.* 2008). For example, the proportions of Turner *et al.*'s outlier regions at FDR = 0.001 that overlap at least one 1-kb window in our 2.5 or 5% F_{ST} tail were 34 and 58%, respectively.

Differentiation across genome annotations: Among CDS, intron, 5′-UTR, 3′-UTR, regulatory, and unannotated parts of the genome, mean F_{ST} was highest for 3′-UTR (Fisher's exact test, P = 0.0007346), in spite of the lower power associated with the small size of the UTR sequence. Moreover, 3′-UTRs were consistently overrepresented in the tail of highly differentiated 1-kb windows (Figure 5). In contrast, coding sequence and introns were consistently underrepresented in the most-differentiated genomic regions. Regions not annotated as either genic or regulatory were also highly enriched in the most-differentiated regions, although less so than 3′-UTRs. Interestingly, regulatory regions and 5′-UTRs were moderately overrepresented in highly differentiated autosomal regions but underrepresented on the *X* chromosome.

To investigate general biological patterns associated with the observed 3′-UTR differentiation, F_{ST} was calculated for each 3′-UTR, which was followed by a Gene Ontology enrichment analysis for the genes associated with the top 1% most-differentiated 3′-UTRs. This analysis revealed no significant enrichments, which was not unexpected given the limited functional annotations associated with most of the genes. However, a number of highly differentiated 3′-UTRs were associated with either transcriptional regulators or genes involved in protein phosphorylation, supporting an important role for regulatory evolution in Queensland *vs.* Tasmania differentiation. Other genes with highly differentiated 3′-UTRs

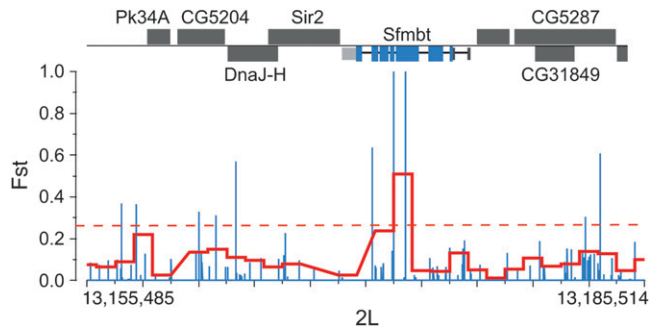

FIGURE 4.—Regions of high population differentiation localize within the *Sfmbt* gene on chromosome *2L*. We plot individual-position F_{ST} (blue) and mean F_{ST} within 1-kb windows (red) across the chromosome. The red dotted line indicates F_{ST} cutoff for the top 2.5% of 1-kb windows. Individual genes are drawn across the top (black); exons are in blue, 3′-UTRs in light gray, and 5′-UTRs in dark gray.

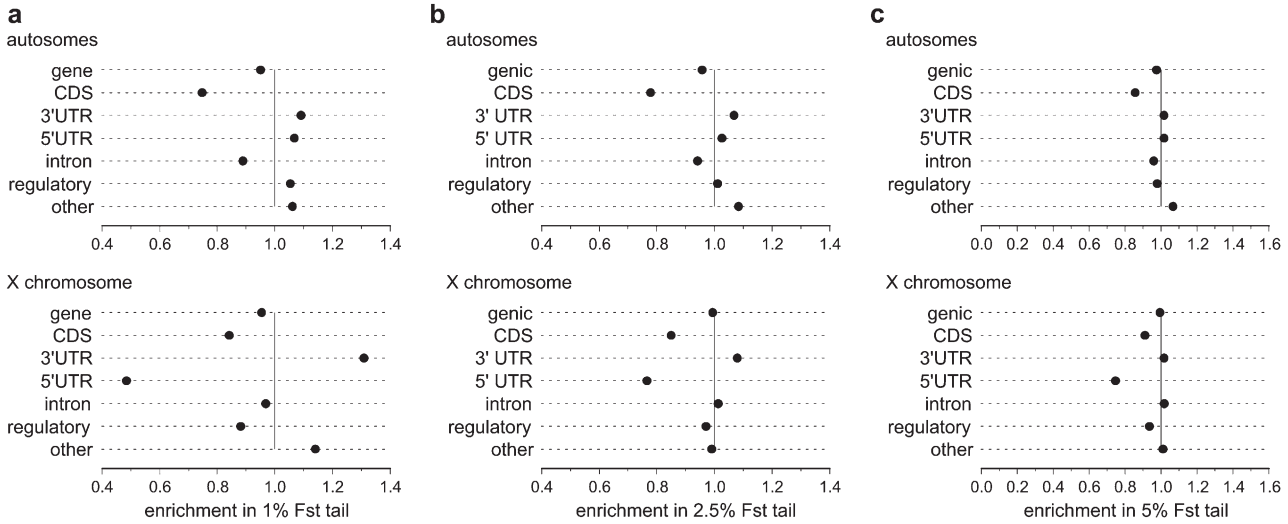

Figure 5.—3′-UTRs and unannotated regions are overrepresented in the most-differentiated genomic regions. We calculated the enrichment for each annotation type in the 1% (a), 2.5% (b), and 5% (c) tail of 1-kb F_{ST} regions, relative to each type's distribution across all 1-kb windows in the normally recombining portion of the genome. Results are shown separately for autosomes and the *X* chromosome. An enrichment score of 1.0 (indicated by a solid vertical line) indicates no enrichment or depletion; values >1 indicate an overabundance of that type in the F_{ST} tail, whereas values <1 indicate underabundance.

code for proteins involved in energy metabolism, development, or seminal fluid (see Table S2).

An example of a gene exhibiting highly localized 3′-UTR differentiation is *Hex-t2,* a testis-specific hexokinase (Duvernell and Eanes 2000). Figure 6 shows that there is a small region of elevated differentiation toward the 3′ end of *Hex-t2*, with peak differentiation occurring in the 3′-UTR. Within this differentiated region are two polymorphic sites in the Queensland population (a U/A polymorphism at position 75 in the UTR and an A/G polymorphism at position 55) that are fixed for the minor allele in Tasmania. Computational prediction of the RNA secondary structure of this 3′-UTR suggests that the Tasmania fixations induce a marked change in RNA secondary structure, consistent with potential functional importance.

Protein-coding differentiation: Despite the fact that many outlier F_{ST} windows fall within exons, coding sequence was not overrepresented in the 1-kb window F_{ST} tail. However, because the windowing analysis does not account for the possibility of different physical scales of selection in DNA sequence space and protein space, alternative methods of characterizing protein differentiation were explored. First, mean F_{ST} for nonsynonymous variants in each gene in the normally recombining portion of the genome was calculated, with the top 1% of individual-gene nonsynonymous F_{ST} considered as coding for highly differentiated proteins. This analysis favors smaller genes/proteins, for which differentiation is likely to be gene/protein-wide. Alternatively, large multidomain proteins might show significant differentiation only in specific functional domains. To investigate this possibility, the PFam database (Finn *et al.* 2010) was used to annotate known functional domains for all *D. melanogaster* genes. Mean nonsynonymous F_{ST} was calculated separately for each domain in a gene, with the maximum domain F_{ST} being recorded for each gene.

Table S3 and Table S4 list the top candidate genes from these analyses, which suggest a number of interesting protein-coding genes for further study. For example, Figure 7a shows elevated differentiation around a fixed amino acid difference at position 47 in the disulfide oxidoreductase gene *Txl*. A threonine residue in Tasmania that is conserved throughout Drosophila has changed to alanine in Queensland, leading to elevated F_{ST} throughout the first exon. The alanine allele has also been observed in African *melanogaster* populations (DPGP.org). This may represent a more unusual case of recent selection in tropical populations (Queensland and Africa) rather than temperate adaptation.

We also observed elevated F_{ST} around a nonsynonymous fixed substitution in *Irc* (Figure 7b), an immune-related catalase required to protect flies from microbial infection (Ha *et al.* 2005a,b). Although the observed V317I substitution in Tasmania is conservative and occurs in a disordered loop region, this position is in direct ligand contact in the protein structure, suggesting a potential functional role in modulating molecular interactions (Figure 7c). Alternatively, these changes could be affecting pre-mRNA processing. The two fixed substitutions in Tasmanian *Irc* are the nonsynonymous V317I change at the 5′ end of exon 6 and a synonymous G → A substitution 11 bases downstream. These changes could be involved in splicing regulation, as RNA secondary structure prediction suggests that they could produce a radical reorganization of pre-mRNA structure (see Figure S4).

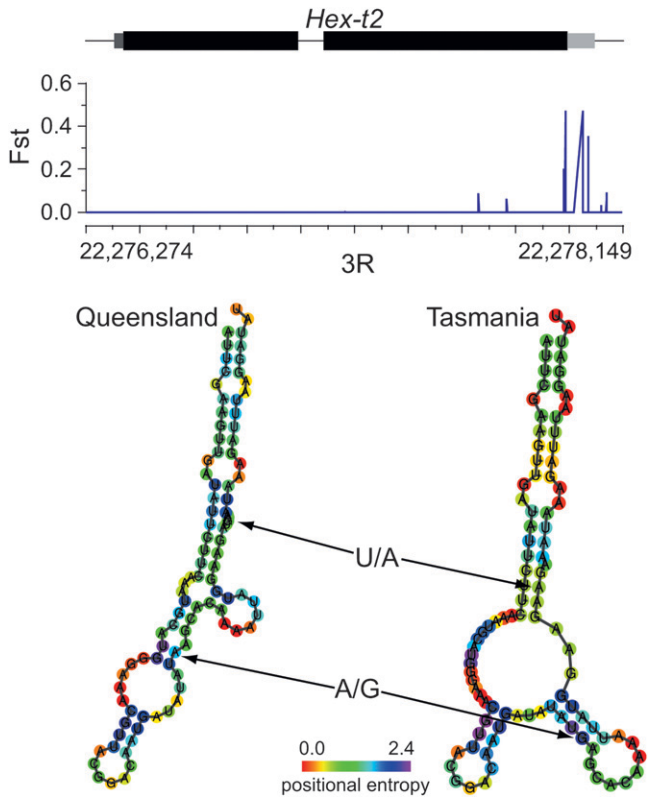

FIGURE 6.—Elevated differentiation between Queensland and Tasmania populations localizes to the 3′-UTR of the *Hex-t2* gene. We plot the F_{ST} of individual genomic positions against the structure of the *Hex-t2* gene. Exons are drawn in black, the 5′-UTR is dark gray, and the 3′-UTR is light gray. The bottom panel shows predicted secondary structures of Queensland and Tasmania 3′-UTR regions. Queensland positions indicated by arrows are polymorphic, with the major allele at left; corresponding positions in Tasmania are fixed for what is the minor allele in Queensland.

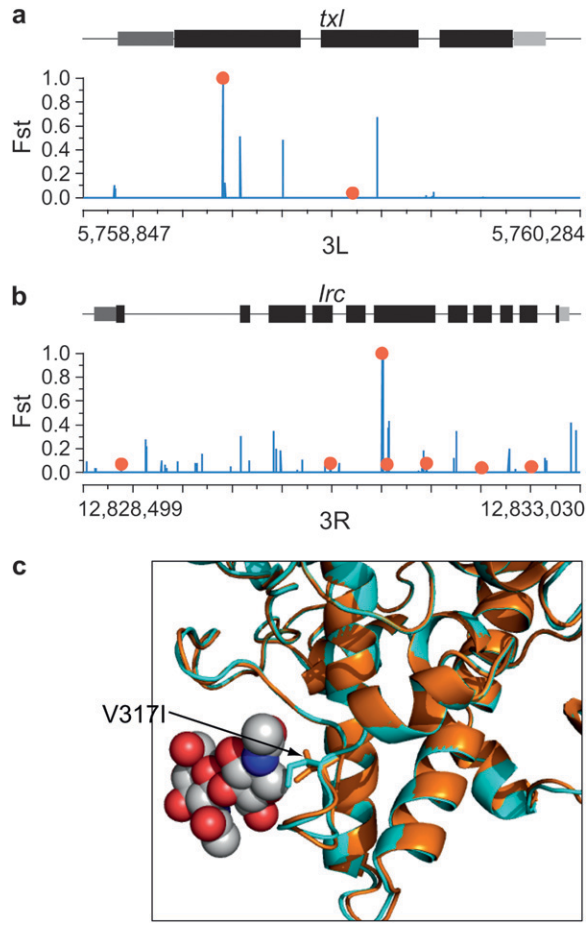

FIGURE 7.—Elevated nonsynonymous F_{ST} in two *melanogaster* protein-coding genes. We plot individual-position F_{ST} along the gene structure. Exons are drawn in black, the 5′-UTR is dark gray, and the 3′-UTR is light gray. Nonsynonymous polymorphisms are shown in red; synonymous and noncoding polymorphisms are shown in blue. (a) A nonsynonymous fixed difference between Queensland and Tasmania is associated with elevated F_{ST} at the *txl* gene. (b) Elevated F_{ST} at a fixed protein-coding change in *Irc*. (c) Structural homology models of Queensland (orange) and Tasmania (turquoise) *Irc*; the V317I substitution is potentially involved in direct ligand interaction.

One of the most differentiated protein domains in the genome is the ligand-binding domain of the *NtR* gene, an extracellular ligand-gated ion channel. Figure 8a shows a large number of polymorphisms across *NtR*, along with a cluster of three amino acid variants in the ligand-binding domain. The most differentiated of these variants is an I/V polymorphism for which the major allele in Queensland (I, frequency = 0.73) is the minor allele in Tasmania (frequency 0.1); F_{ST} for this site is 0.51. The remaining amino acid polymorphisms in this domain are an L/F polymorphism ($F_{ST} = 0.14$) and an E/D polymorphism ($F_{ST} = 0.19$). While L is the major allele in both populations at the first position, the E/D Queensland polymorphism is fixed for D in Tasmania. Structural homology modeling suggests that this E/D polymorphism occurs in the main immunogenic region (MIR) of the protein (Figure 8b). This region constitutes a loop sandwiched between β2 and β3 that binds autoimmune antibodies in myasthenia gravis patients in the homologous human muscle acetylcholine receptor (TSOULOUFIS *et al.* 2000; DELLISANTI *et al.* 2007). The fact that the I/V polymorphism is found in close proximity to this region suggests the possibility that differentiation at *NtR* could affect interactions with other molecules, possibly those relating to the immune system.

Biological patterns underlying genic differentiation: The extensive genetic interactions and pleiotropic effects of laboratory mutations in Drosophila genes make it challenging to reliably infer from differentiated genes the phenotypes that may be targets of selection. Nevertheless, the small physical scale of differentiation makes it worthwhile to explore general patterns in the data as a means of generating hypotheses regarding pathways and phenotypes that might experience spatially varying selection in Australian *melanogaster* populations. Our approach was to test for enrichment of GO terms among the genes that overlapped a 1-kb window in the upper 2.5% tail of the distribution, which

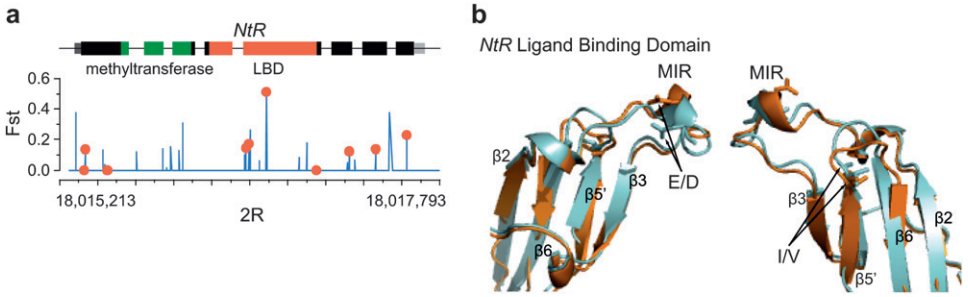

FIGURE 8.—Elevated nonsynonymous differentiation in *NtR* localizes to the major immunogenic region (MIR) of the ligand-binding domain (LBD). (a) We plot positional F_{ST} across gene structure, with exons drawn in black, 5′-UTR in dark gray, and 3′-UTR in light gray; methyltransferase and ligand-binding domains are indicated by green and red, respectively. Nonsynonymous polymorphisms are shown by red circles. (b) We plot highly differentiated E/D and I/V polymorphisms on the predicted 3D structure of the *NtR* LBD. In both cases, the major allele in Queensland (E, I) is shown in orange, and the major allele in Tasmania (D, V) is shown in turquoise.

corresponds to $F_{ST} > 0.27$. These analyses were supplemented by inspection of genetic interactions annotated in FlyBase. We also point to plausible candidates in the 5% tail where appropriate.

Several high-F_{ST} windows overlapped genes functioning in central Drosophila signaling pathways, including the *JAK-STAT* pathway, the *torso* pathway, the *EGFR* pathway, and the TGF-β pathway. In the *JAK-STAT* pathway the ligand *upd2* (1-kb $F_{ST} = 0.70$) and *STAT* (*Stat92E*, 1-kb $F_{ST} = 0.32$) both showed elevated F_{ST}, as did *CycE* (1-kb $F_{ST} = 0.25$) and *Ptp61F* (1-kb $F_{ST} = 0.28$), which regulate that pathway. Other modifiers of *JAK-STAT* signaling that overlapped high-F_{ST} windows included *crb* (1-kb $F_{ST} = 0.35$), *tkv* (1-kb $F_{ST} = 0.39$), *Mad* (1-kb $F_{ST} = 0.35$), and *Stam* (1-kb $F_{ST} = 0.33$). Highly differentiated genes in the *torso* signaling pathway (which regulates several processes, including metamorphosis and body size) included *tup* (1-kb $F_{ST} = 0.41$), *Gap1* (1-kb $F_{ST} = 0.26$), *pnt* (1-kb $F_{ST} = 0.60$), *tld* (1-kb $F_{ST} = 0.25$), and *csw* (1-kb $F_{ST} = 0.26$). Differentiated genes in the *EGFR* signaling pathway included *vn* (1-kb $F_{ST} = 0.27$), *argos* (1-kb $F_{ST} = 0.23$), *sprouty* (1-kb $F_{ST} = 0.29$), *Star* (1-kb $F_{ST} = 0.29$), and *ed* (1-kb $F_{ST} = 0.30$). Genes in the TGF-β pathway were also overrepresented among high-F_{ST} windows and included *dally* (1-kb $F_{ST} = 0.39$), *Mad*, and *tkv* (1-kb $F_{ST} = 0.39$). The gene *dpp*, which is centrally located in this pathway, also contained a region of high differentiation (1-kb $F_{ST} = 0.24$). Finally, the hypothesis that ecdysone signaling experiences spatially varying selection is supported by highly differentiated windows overlapping the ecdysone receptor, *EcR* (1-kb $F_{ST} = 0.25$); the eclosion hormone gene *Eh* (1-kb $F_{ST} = 0.33$); *Moses* (1-kb $F_{ST} = 0.41$); *taiman* (1-kb $F_{ST} = 0.37$); and the ecdysone-induced protein-coding genes *Eip63E* (1-kb $F_{ST} = 0.33$), *Eip74EF* (1-kb $F_{ST} = 0.31$), *Eip75B* (1-kb $F_{ST} = 0.30$), and *Eip93F* (1-kb $F_{ST} = 0.44$). It is worth noting that substantial crosstalk exists between some of these pathways and that other genes associated with key pathways such as *Notch* show evidence of differentiation in our data.

These results support the existence of pervasive spatially varying selection acting at key genes throughout multiple Drosophila signaling pathways. It is highly plausible that several candidates influence clinal variation in body size, metabolism, and additional important life history traits (see Table S5 for a complete list of enriched GO terms). Many genes implicated in body-size variation were highly differentiated, including *InR* [1-kb $F_{ST} = 0.26$ (PAABY *et al.* 2010)], *dally* (1-kb $F_{ST} = 0.39$), *Orct2*, and *Pi3K21B* at the tip of 2L, which contains a highly differentiated 1-kb window ($F_{ST} = 0.28$) but was not included in most of our analyses because of its location at the distal end of the chromosome arm. Interestingly, many body-size candidate genes revealed by our analysis are located on chromosome arm *3R*, which is consistent with previous genetic analyses showing that most of the body-size variation associated with the Australian cline is inseparable from *In(3R)P* in mapping crosses (RAKO *et al.* 2006, 2007). Our data—including evidence of extensive recombination between standard and *In(3R)P* arrangements—suggest that the differentiated genes that are located on *3R* are particularly promising targets for investigating the genetic basis of body-size variation in *D. melanogaster*.

A large number of GO terms related to developmental processes are enriched for F_{ST} outliers. The associated genes contribute to many phenotypes, including external morphology (*e.g.*, wing and eye), nervous system development, ovarian follicle development, larval development, and embryonic development. The *Toll* signaling pathway, which contains a number of immune system genes, is enriched. The immunity gene *sick* is also in the 5% tail of F_{ST} windows. Olfactory behavior and olfactory learning are enriched in 1-kb outlier tails. In addition, a number of F_{ST}-outlier nonsynonymous SNPs not located in outlier windows are found in olfactory or gustatory receptors or odorant-binding proteins. Several ionotropic receptors, a new class of odorant receptors, appear in the 5% F_{ST} tail of 1-kb windows. It is interesting to note the evidence that thermal stress disrupts odor learning in flies (WANG *et al.* 2007) via developmental effects on the mushroom body, in light of the observation that "mushroom body development" is among the enriched GO terms in our analysis. A number of ion channel-related genes appear among the outlier 1-kb windows, leading to enrichment of GO categories:

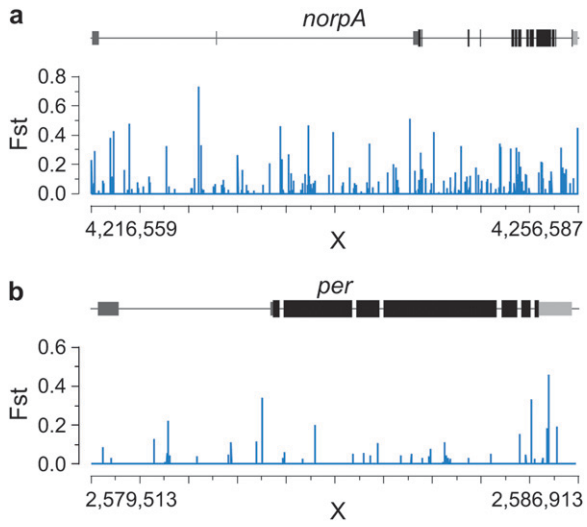

FIGURE 9.—Coordinated differentiation in *norpA* (a) and the 3′-UTR of *per* (b), a known target of *norpA* splicing regulation. We plot individual-position F_{ST} along the gene structure. Exons are drawn in black, the 5′-UTR is dark gray, and the 3′-UTR is light gray.

calcium-, potassium-, and sodium-ion transport. "Calcium ion binding" is the second most significantly enriched molecular function and includes several Cadherins as well as Calmodulin. Selection associated with variation in the visual environment between Queensland and Tasmania is suggested by the enrichment of GO terms such as "phototransduction."

Although circadian rhythm genes are not overrepresented among the F_{ST} outliers, several genes relating to circadian biology are found among the most differentiated 1-kb windows. The cryptochrome gene, which regulates circadian rhythm, is highly differentiated ($F_{ST} = 0.30$), as are *couch potato* ($F_{ST} = 0.23$) and *timeless* ($F_{ST} = 0.20$), which have already been implicated in spatially varying selection in *D. melanogaster* (SANDRELLI *et al.* 2007; TAUBER *et al.* 2007; SCHMIDT *et al.* 2008). Another interesting candidate is *norpA*, a phospholipase C gene required for thermal synchronization of the circadian clock (GLASER and STANEWSKY 2005). This gene is in the 2.5% F_{ST} tail and highly differentiated across its entire length (see Figure 9a). Four of its seven interacting partners annotated in FlyBase are also in the 2.5% tail (see Table S7). Additionally, *norpA* is known to regulate splicing in the 3′-UTR of *per*, a central circadian-clock gene in Drosophila (COLLINS *et al.* 2004; MAJERCAK *et al.* 2004) that shows a highly localzed 3′-UTR elevation in F_{ST} in our data (Figure 9b). Together, these results strongly suggest a cluster of correlated differentiation occurring across several genes at the interface between thermal and light entrainment of the circadian clock.

Finally, transcription and chromatin regulation appear to be under widespread selection, as seven related biological process GO terms are enriched among the F_{ST} outlier windows. Additionally, "transcription factor" is the second most significantly enriched GO molecular function term. Particularly interesting differentiated genes include *Trl*, *HDAC4*, *additional sex combs*, *Enhancer of polycomb*, *histoneacetlytransferase Tip60*, *Ino80*, *JIL-1*, *14-3-3ε*, and *Sfmbt*.

Copy-number variation: Differences in copy number between Queensland and Tasmania were investigated using an outlier approach analogous to that used for F_{ST}. The normalized ratio of Queensland/Tasmania coverage for 1-kb nonoverlapping windows was calculated across the genome (see MATERIALS AND METHODS), with the top 1% most-extreme estimates considered highly differentiated regions. Note that frequency variation and ploidy-level variation are confounded in this analysis. Relative to the genome-wide average of copy-number differentiation, slightly more than half (55%) of the 1-kb windows had more coverage in the Queensland population. However, significantly more (62.5%) of the highly differentiated windows showed increased copy number in the Tasmania population ($P = 2.2 \times 10^{-16}$), suggesting that duplication events could be important for local adaptation in Tasmania.

The largest region exhibiting significant copy-number variation (CNV) is a 107-kb region of chromosome *3R* (Figure 10), which spans a small number of protein-coding genes including the last few exons of *timeout* and the entire *Ace* gene. *Ace* codes for an acetylcholinesterase associated with pesticide resistance (MENOZZI *et al.* 2004), which was previously identified as a differentiated CNV between these populations (TURNER *et al.* 2008). Interestingly, *Ace* expression has been shown to vary over the circadian cycle (HOOVEN *et al.* 2009), and acetylcholinesterase levels are highly correlated with pesticide resistance (CHARPENTIER and FOURNIER 2001).

Gene Ontology enrichment analysis of genes found in highly differentiated CNV regions revealed categories similar to those observed for our F_{ST} enrichment analysis (see Table S6), including transcription factors and ion-channel genes. Across both GO-enrichment analyses, 185 unique GO terms were enriched, 66 of which (36%) were found in both analyses. Interestingly, despite the large degree of overlap between GO enrichment terms in the F_{ST} and CNV analyses, the specific genes associated with each enriched GO category did not overlap to a large degree. Of the 719 genes in the copy-number 1% outlier set and the 551 genes in the corresponding F_{ST} outlier set, only 72 (6%) were found in both (as expected given the upper bound of coverage included in the F_{ST} analysis). This suggests the possibility that selection may often result in recruitment of alleles resulting from both nucleotide and copy-number differences. Several terms enriched in the CNV GO analysis did not appear in the F_{ST} GO enrichment, including "circadian rhythm," "sex determination," "courtship and mating behavior," "female meiosis chromosome segregation," and "chorion-containing eggshell formation" (which was also detected by TURNER *et al.* 2008).

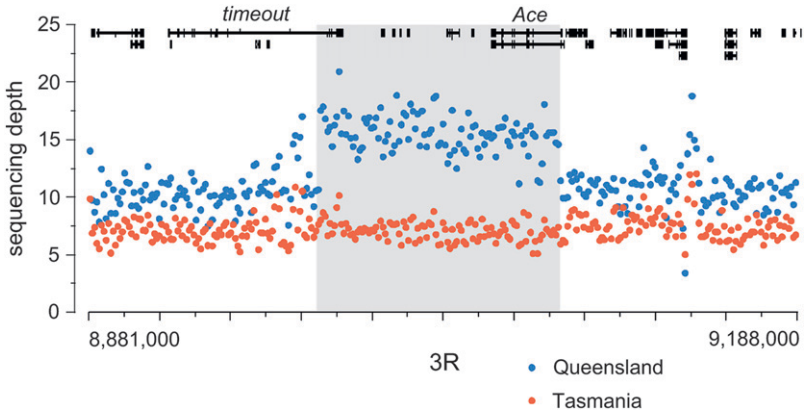

FIGURE 10.—A large region of increased copy number in Queensland occurs on chromosome *3R*. We plot the average number of sequence reads for each 1-kb window across this region, both for the Queensland (blue) and for the Tasmania (red) populations. Genes in this region are drawn across the top. The gray box indicates the inferred region of increased copy number in Queensland.

DISCUSSION

A large body of evidence supports the idea that much of the phenotypic and genetic differentiation along the Australian *D. melanogaster* latitudinal cline is driven by spatially varying selection (OAKESHOTT *et al.* 1981, 1983; SINGH and RHOMBERG 1987; SINGH 1989; SINGH and LONG 1992; GOCKEL *et al.* 2001; KENNINGTON *et al.* 2003; HOFFMANN and WEEKS 2007). Here we have presented the first genome-sequence-based analysis of population differentiation associated with this cline. Although our analysis included only populations from each end of the cline, it is likely that the set of highly differentiated genomic regions between these cline endpoints is considerably enriched for targets of spatially varying selection. Indeed, the fact that the most highly differentiated genomic regions show much more negative Fay and Wu's *H* estimates in Tasmania is consistent with the hypothesis that the observed differentiation is associated with recent strong selection in temperate populations (SEZGIN *et al.* 2004). The dramatic enrichment of several GO terms among the genes overlapping differentiated regions also supports the notion that selection plays a major role, because it is difficult to envision a neutral demographic process that could result in such enrichment patterns.

Two main lines of evidence support the proposition that gene regulation is an important target of spatially varying selection in these populations. First, 3′-UTRs and unannotated sequence are the most overrepresented sequence classes among the outlier 1-kb F_{ST} windows. 3′-UTRs, which exhibit the strongest enrichment in our analysis, play an important role in gene regulation (LAI 2002; KUERSTEN and GOODWIN 2003; DE MOOR *et al.* 2005; STARK *et al.* 2005; CHATTERJEE and PAL 2009; MANGONE *et al.* 2010). Recent studies have found substantial *cis*-acting effects on regulatory variation in Drosophila (HUGHES *et al.* 2006; LAWNICZAK *et al.* 2008; LEMOS *et al.* 2008; GRAZE *et al.* 2009; MCMANUS *et al.* 2010); our results raise the intriguing possibility that variation in 3′-UTRs may make a significant contribution to adaptive *cis*-acting regulatory variation. The overrepresentation of noncoding DNA among F_{ST} outlier windows is consistent with previous population genetic results supporting the importance of noncoding sequence for adaptive divergence over longer timescales in *D. melanogaster* (ANDOLFATTO 2005). It will be interesting to investigate these currently unannotated regions in the context of ongoing efforts to improve the annotation of the *D. melanogaster* genome (CELNIKER *et al.* 2009). The second line of evidence supporting the importance of selection on gene regulation along the cline is the finding that transcription- and chromatin-related genes are among the most differentiated in the genome, which is consistent with previous analyses of these populations (LEVINE and BEGUN 2008; TURNER *et al.* 2008) and with genomic inferences on the importance of recurrent directional selection on proteins regulating chromatin and transcription in *D. simulans* (BEGUN *et al.* 2007).

Although the protein-coding sequence was underrepresented among the most extremely differentiated 1-kb windows, one should not conclude that amino acid variants are unimportant for selection along the cline, as a large number of outlier windows overlap coding sequence. It is interesting to consider possible population-genetic explanations for why CDS is underrepresented. The timescale of differentiation between Queensland and Tasmanian populations is very small, perhaps on the order of 1000 generations (HOFFMANN and WEEKS 2007). Because the mutation rate per base pair is small, much of the selective response during the initial colonization of Australia was likely the result of frequency changes of alleles already segregating in ancestral populations rather than from invasion into the populations of new mutations that occurred subsequent to colonization. Whole-genome surveys of polymorphism in Drosophila suggest that nonsynonymous sites are severalfold less polymorphic than synonymous or noncoding sites (*e.g.*, BEGUN *et al.* 2007; SACKTON *et al.* 2009). Thus, on a per-site basis compared to noncoding variants, amino acid variants are considerably less available to selection on standing variation following a radical change of the environment. The physical scale

of differentiation predicted under the selection-on-standing-variation model depends on the amount of linkage disequilibrium associated with the site destined to experience selection after the environment changes. Surveys of linkage disequilibrium in normally recombining regions from large samples of cosmopolitan *D. melanogaster* consistently find that sites in strong linkage disequilibrium tend to be within 2 kb of each other (Miyashita and Langley 1988; Palsson *et al.* 2004; Macdonald *et al.* 2005). This is consistent with the scale of geographic differentiation observed in our data and with the hypothesis that much of the observed differentiation between temperate and tropical populations is the result of recent strong selection on standing variants. Genomic data on the frequency distribution of variation and the scale of linkage disequilbrium from populations along the Australian cline and from African and European populations should provide the resources necessary for addressing issues relating to the geographic origins, frequencies, and fitnesses of variants experiencing selection in Australia.

One of the general findings from our analysis is that many genes and pathways centrally important to Drosophila biology appear to experience spatially varying selection. The fact that laboratory mutations in these genes and pathways tend to be highly pleiotropic is, in the conventional thinking, associated with reduced mutation rate to beneficial alleles. It is important to realize, however, that it is the individual mutation—rather than the gene—that is more or less pleiotropic. The distribution of pleiotropic effects of natural variants is likely to be quite different and dramatically smaller than those of laboratory mutations. Moreover, the large population sizes of Drosophila suggest that drift may be relatively unimportant and that variants that reach appreciable frequencies may have special genetic and population-genetic properties. Thus, the candidate variants identified here may have very small pleiotropic effects, in spite of the fundamental biological roles of the corresponding genes. Alternatively, natural alleles that were pleiotropic along the axes favored by correlated natural selection would be strongly favored, and these too could constitute a considerable fraction of the variants in fundamental signaling pathways that show differentiation between these populations.

The genomic results regarding the dramatic biological differences between these fly populations raise the obvious question—unanswerable with these data—as to the phenotypic and fitness effects of the selected mutations and how the distribution of such effects may vary across biological functions and positions in genetic pathways. For example, one class of selected mutations may contribute to phenotypic differences between temperate and tropical flies, while a second—potentially larger—class exhibiting genotype × environment interactions may exhibit latitudinal clines, because different genotypes are required to produce a single optimal phenotype in different environments (*e.g.*, Levine *et al.* 2011). Larger genomic data sets and functional analyses should produce much sharper inferences regarding the specific polymorphisms, pathways, and biological functions that have diverged under selection between temperate and tropical populations and further reveal the genetic and population-genetic principles of adaptation in this model species.

We thank Ary Hoffmann for generously sharing flies and A. Hoffmann and P. Schmidt for their thoughts on clinal variation in Drosophila. We thank Michael Nachman, J. Anderson, and two anonymous reviewers for comments that improved this manuscript. We also gratefully acknowledge Charis Cardeno, Kristian Stevens, Melissa Eckert, and Thaddeus Seher for technical assistance and Phil Nista for early contributions to the analysis. This work was funded by National Institutes of Health grants GM071926 and GM084056 (to D.J.B.), by the Drosophila Population Genomics Project (Chuck Langley, PI), and by Dartmouth College and the Neukom Institute (A.D.K.).

LITERATURE CITED

Achaz, G., 2009 Frequency spectrum neutrality tests: one for all and all for one. Genetics **183:** 249–258.

Adams, M. D., S. E. Celniker, R. A. Holt, C. A. Evans, J. D. Gocayne *et al.*, 2000 The genome sequence of Drosophila melanogaster. Science **287**(5461): 2185–2195.

Aguade, M., N. Miyashita and C. H. Langley, 1989 Reduced variation in the yellow-achaete-scute region in natural populations of *Drosophila melanogaster*. Genetics **122:** 607–615.

Akey, J. M., G. Zhang, K. Zhang, L. Jin and M. D. Shriver, 2002 Interrogating a high-density SNP map for signatures of natural selection. Genome Res. **12**(12): 1805–1814.

Anderson, A., A. Hoffmann, S. Mckechnie, P. Umina and A. Weeks, 2005 The latitudinal cline in the In (3 R) Payne inversion polymorphism has shifted in the last 20 years in Australian Drosophila melanogaster populations. Mol. Ecol. **14**(3): 851–858.

Andolfatto, P., 2005 Adaptive evolution of non-coding DNA in Drosophila. Nature **437**(7062): 1149–1152.

Ashburner, M., C. A. Ball, J. A. Blake, D. Botstein, H. Butler *et al.*, 2000 Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat. Genet. **25**(1): 25–29.

Beaumont, M., and R. Nichols, 1996 Evaluating loci for use in the genetic analysis of population structure. Proc. R. Soc. B Biol. Sci. **263**(1377): 1619–1626.

Beaumont, M. A., and D. J. Balding, 2004 Identifying adaptive genetic divergence among populations from genome scans. Mol. Ecol. **13**(4): 969–980.

Begun, D., A. Holloway, K. Stevens, L. Hillier, Y. Poh *et al.*, 2007 Population genomics: whole-genome analysis of polymorphism and divergence in Drosophila simulans. PLoS Biol. **5**(11): e310.

Begun, D. J., and C. F. Aquadro, 1995 Evolution at the tip and base of the X chromosome in an African population of Drosophila melanogaster. Mol. Biol. Evol. **12**(3): 382–390.

Bentley, D. R., S. Balasubramanian, H. P. Swerdlow, G. P. Smith, J. Milton *et al.*, 2008 Accurate whole human genome sequencing using reversible terminator chemistry. Nature **456**(7218): 53–59.

Berman, H. M., J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat *et al.*, 2000 The Protein Data Bank. Nucleic Acids Res. **28**(1): 235–242.

Blanchette, M., W. J. Kent, C. Riemer, L. Elnitski, A. F. A. Smit *et al.*, 2004 Aligning multiple genomic sequences with the threaded blockset aligner. Genome Res. **14**(4): 708–715.

Celniker, S. E., L. A. Dillon, M. B. Gerstein, K. C. Gunsalus, S. Henikoff *et al.*, 2009 Unlocking the secrets of the genome. Nature **459**(7249): 927–930.

Chan, Y., M. Marks, F. Jones, G. Villarreal, Jr, M. Shapiro *et al.*, 2010 Adaptive evolution of pelvic reduction in sticklebacks by recurrent deletion of a Pitx1 enhancer. Science **327**(5963): 302.

Charpentier, A., and D. Fournier, 2001 Levels of total acetylcholinesterase in Drosophila melanogaster in relation to insecticide resistance. Pestic. Biochem. Physiol. **70**(2): 100–107.

Chatterjee, S., and J. K. Pal, 2009 Role of 5′- and 3′-untranslated regions of mRNAs in human diseases. Biol. Cell **101**(5): 251–262.

Coggill, P., R. D. Finn and A. Bateman, 2008 Identifying protein domains with the Pfam database. Curr. Protoc. Bioinformatics, Chap. 2.

Collins, B. H., E. Rosato and C. P. Kyriacou, 2004 Seasonal behavior in Drosophila melanogaster requires the photoreceptors, the circadian clock, and phospholipase C. Proc. Natl. Acad. Sci. USA **101**(7): 1945–1950.

Colosimo, P., C. Peichel, K. Nereng, B. Blackman, M. Shapiro *et al.*, 2004 The genetic architecture of parallel armor plate reduction in threespine sticklebacks. PLoS Biol. **2**(5): 635–641.

Costa, R., A. A. Peixoto, G. Barbujani and C. P. Kyriacou, 1992 A latitudinal cline in a Drosophila clock gene. Proc. Biol. Sci. **250**(1327): 43–49.

Daborn, P. J., J. L. Yen, M. R. Bogwitz, G. Le Goff, E. Feil *et al.*, 2002 A single p450 allele associated with insecticide resistance in Drosophila. Science **297**(5590): 2253–2256.

de Bono, M., and C. I. Bargmann, 1998 Natural variation in a neuropeptide Y receptor homolog modifies social behavior and food response in C. elegans. Cell **94**(5): 679–689.

de Moor, C. H., H. Meijer and S. Lissenden, 2005 Mechanisms of translational control by the 3′ UTR in development and differentiation. Semin. Cell Dev. Biol. **16**(1): 49–58.

Dellisanti, C. D., Y. Yao, J. C. Stroud, Z. Z. Wang and L. Chen, 2007 Crystal structure of the extracellular domain of nAChR alpha1 bound to alpha-bungarotoxin at 1.94 A resolution. Nat. Neurosci. **10**(8): 953–962.

Drosophila 12 Genomes Consortium, 2007 Evolution of genes and genomes on the Drosophila phylogeny. Nature **450**(7167): 203–218.

Duvernell, D. D., and W. F. Eanes, 2000 Contrasting molecular population genetics of four hexokinases in *Drosophila melanogaster, D. simulans* and *D. yakuba*. Genetics **156:** 1191–1201.

Duvernell, D. D., P. S. Schmidt and W. F. Eanes, 2003 Clines and adaptive evolution in the methuselah gene region in Drosophila melanogaster. Mol. Ecol. **12**(5): 1277–1285.

Eramian, D., N. Eswar, M. Y. Shen and A. Sali, 2008 How well can the accuracy of comparative protein structure models be predicted? Protein Sci. **17**(11): 1881–1893.

Eswar, N., D. Eramian, B. Webb, M. Y. Shen and A. Sali, 2008 Protein structure modeling with MODELLER. Methods Mol. Biol. **426:** 145–159.

Ewens, W. J., 2004 *Mathematical Population Genetics*, Ed. 2, Vol. 27. Springer, New York.

Fay, J. C., and C. I. Wu, 2000 Hitchhiking under positive Darwinian selection. Genetics **155:** 1405–1413.

Finn, R. D., J. Mistry, J. Tate, P. Coggill, A. Heger *et al.*, 2010 The Pfam protein families database. Nucleic Acids Res. **38**(Database issue): 211–222.

Frydenberg, J., A. A. Hoffmann and V. Loeschcke, 2003 DNA sequence variation and latitudinal associations in hsp23, hsp26 and hsp27 from natural populations of Drosophila melanogaster. Mol. Ecol. **12**(8): 2025–2032.

Fu, Y. X., 1995 Statistical properties of segregating sites. Theor. Popul. Biol. **48:** 172–197.

Futschik, A., and C. Schlotterer, 2010 The next generation of molecular markers from massively parallel sequencing of pooled DNA samples. Genetics **186:** 207–218.

Glaser, F. T., and R. Stanewsky, 2005 Temperature synchronization of the Drosophila circadian clock. Curr. Biol. **15**(15): 1352–1363.

Gockel, J., W. J. Kennington, A. Hoffmann, D. B. Goldstein and L. Partridge, 2001 Nonclinality of molecular variation implicates selection in maintaining a morphological cline of Drosophila melanogaster. Genetics **158:** 319–323.

Graze, R. M., L. M. McIntyre, B. J. Main, M. L. Wayne and S. V. Nuzhdin, 2009 Regulatory divergence in *Drosophila melanogaster* and *D. simulans*, a genomewide analysis of allele-specific expression. Genetics **183:** 547–561.

Griffith, O. L., S. B. Montgomery, B. Bernier, B. Chu, K. Kasaian *et al.*, 2008 ORegAnno: an open-access community-driven resource for regulatory annotation. Nucleic Acids Res. **36**(Database issue): D107–D113.

Grimm, C., R. Matos, N. Ly-Hartig, U. Steuerwald, D. Lindner *et al.*, 2009 Molecular recognition of histone lysine methylation by the Polycomb group repressor dSfmbt. EMBO J. **28**(13): 1965–1977.

Ha, E. M., C. T. Oh, Y. S. Bae and W. J. Lee, 2005a A direct role for dual oxidase in Drosophila gut immunity. Science **310**(5749): 847–850.

Ha, E. M., C. T. Oh, J. H. Ryu, Y. S. Bae, S. W. Kang *et al.*, 2005b An antioxidant system required for host protection against gut infection in Drosophila. Dev. Cell **8**(1): 125–132.

Hasegawa, M., H. Kishino and T. Yano, 1985 Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J. Mol. Evol. **22**(2): 160–174.

Hinrichs, A. S., D. Karolchik, R. Baertsch, G. P. Barber, G. Bejerano *et al.*, 2006 The UCSC Genome Browser Database: update 2006. Nucleic Acids Res. **34**(Database issue): D590–D598.

Hofacker, I. L., 2003 Vienna RNA secondary structure server. Nucleic Acids Res. **31**(13): 3429–3431.

Hoffmann, A. A., and A. R. Weeks, 2007 Climatic selection on genes and traits after a 100 year-old invasion: a critical look at the temperate-tropical clines in Drosophila melanogaster from eastern Australia. Genetica **129**(2): 133–147.

Hooven, L. A., K. A. Sherman, S. Butcher and J. M. Giebultowicz, 2009 Does the clock make the poison? Circadian variation in response to pesticides. PLoS One **4**(7): e6469.

Hughes, K. A., J. F. Ayroles, M. M. Reedy, J. M. Drnevich, K. C. Rowe *et al.*, 2006 Segregating variation in the transcriptome: *cis* regulation and additivity of effects. Genetics **173:** 1347–1355.

Katoh, K., and H. Toh, 2008 Recent developments in the MAFFT multiple sequence alignment program. Brief. Bioinformatics **9**(4): 286–298.

Katoh, K., K. Misawa, K. Kuma and T. Miyata, 2002 MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. **30**(14): 3059–3066.

Kennington, W. J., J. Gockel and L. Partridge, 2003 Testing for asymmetrical gene flow in a Drosophila melanogaster body-size cline. Genetics **165:** 667–673.

Knibb, W. R., 1982 Chromosome inversion polymorphisms in Drosophila melanogaster. II. Geographic clines and climatic associations in Australasia, North America and Asia. Genetica **53**(3): 213–221.

Knibb, W. R., J. G. Oakeshott and J. B. Gibson, 1981 Chromosome inversion polymorphisms in Drosophila melanogaster. I. Latitudinal clines and associations between inversions in Australasian populations. Genetics **98:** 833–847.

Kuersten, S., and E. B. Goodwin, 2003 The power of the 3′ UTR: translational control and development. Nat. Rev. Genet. **4**(8): 626–637.

Lai, E. C., 2002 Micro RNAs are complementary to 3′ UTR sequence motifs that mediate negative post-transcriptional regulation. Nat. Genet. **30**(4): 363–364.

Langley, C. H., B. P. Lazzaro, W. Phillips, E. Heikkinen and J. M. Braverman, 2000 Linkage disequilibria and the site frequency spectra in the su(s) and su(w(a)) regions of the *Drosophila melanogaster* X chromosome. Genetics **156:** 1837–1852.

Lawniczak, M. K., A. K. Holloway, D. J. Begun and C. D. Jones, 2008 Genomic analysis of the relationship between gene expression variation and DNA polymorphism in Drosophila simulans. Genome Biol. **9**(8):R125.

Lemos, B., L. O. Araripe, P. Fontanillas and D. L. Hartl, 2008 Dominance and the evolutionary accumulation of cis- and trans-effects on gene expression. Proc. Natl. Acad. Sci. USA **105**(38): 14471–14476.

Levine, M. T., and D. J. Begun, 2008 Evidence of spatially varying selection acting on four chromatin-remodeling loci in *Drosophila melanogaster*. Genetics **179:** 475–485.

Levine, M. T., M. Eckert and D. J. Begun, 2011 Whole-genome expression plasticity across tropical and temperate *Drosophila melanogaster* populations from eastern Australia. Mol. Biol. Evol. **28(1):** 249–256.

Lewontin, R. C., 1974 *The Genetic Basis of Evolutionary Change*, Vol. 25. Columbia University Press, New York.

Lewontin, R. C., and J. Krakauer, 1975 Letters to the editors: testing the heterogeneity of F values. Genetics **80:** 397–398.

Li, H., J. Ruan and R. Durbin, 2008 Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res. **18**(11): 1851–1858.

MacArthur, S., X.-Y. Li, J. Li, J. B. Brown, H. C. Chu *et al.*, 2009 Developmental roles of 21 Drosophila transcription factors are determined by quantitative differences in binding to an overlapping set of thousands of genomic regions. Genome Biol. **10**(7): R80.

Macdonald, S. J., T. Pastinen and A. D. Long, 2005 The effect of polymorphisms in the enhancer of split gene complex on bristle number variation in a large wild-caught cohort of *Drosophila melanogaster*. Genetics **171:** 1741–1756.

Majercak, J., W. F. Chen and I. Edery, 2004 Splicing of the period gene 3′-terminal intron is regulated by light, circadian clock factors, and phospholipase C. Mol. Cell. Biol. **24**(8): 3359–3372.

Mangone, M., A. P. Manoharan, D. Thierry-Mieg, J. Thierry-Mieg, T. Han *et al.*, 2010 The landscape of C. elegans 3′UTRs. Science **329**(5990): 432–435.

Maruyama, T., 1970 On the rate of decrease of heterozygosity in circular stepping stone models of populations* 1. Theor. Popul. Biol. **1**(1): 101–119.

McColl, G., and S. W. McKechnie, 1999 The Drosophila heat shock hsr-omega gene: an allele frequency cline detected by quantitative PCR. Mol. Biol. Evol. **16**(11): 1568–1574.

McManus, C. J., J. D. Coolon, M. O. Duff, J. Eipper-Mains, B. R. Graveley *et al.*, 2010 Regulatory divergence in Drosophila revealed by mRNA-seq. Genome Res. **20**(6): 816–825.

Menozzi, P., M. A. Shi, A. Lougarre, Z. H. Tang and D. Fournier, 2004 Mutations of acetylcholinesterase which confer insecticide resistance in Drosophila melanogaster populations. BMC Evol. Biol. **4**: 4.

Miller, C., S. Beleza, A. Pollen, D. Schluter, R. Kittles *et al.*, 2007 cis-regulatory changes in Kit ligand expression and parallel evolution of pigmentation in sticklebacks and humans. Cell **131**(6): 1179–1189.

Miyashita, N., and C. H. Langley, 1988 Molecular and phenotypic variation of the white locus region in *Drosophila melanogaster*. Genetics **120:** 199–212.

Nachman, M., H. Hoekstra and S. D'Agostino, 2003 The genetic basis of adaptive melanism in pocket mice. Proc. Natl. Acad. Sci. USA **100**(9): 5268.

Nei, M., 1987 *Molecular Evolutionary Genetics*. Columbia University Press, New York.

Oakeshott, J. G., G. K. Chambers, J. B. Gibson and D. A. Willcocks, 1981 Latitudinal relationships of esterase-6 and phosphoglucomutase gene frequencies in Drosophila melanogaster. Heredity **47**(Pt. 3): 385–396.

Oakeshott, J. G., G. K. Chambers, J. B. Gibson, W. F. Eanes and D. A. Willcocks, 1983 Geographic variation in G6pd and Pgd allele frequencies in Drosophila melanogaster. Heredity **50**(Pt. 1): 67–72.

Osborne, K., A. Robichon, E. Burgess, S. Butland, R. Shaw *et al.*, 1997 Natural behavior polymorphism due to a cGMP-dependent protein kinase of Drosophila. Science **277**(5327): 834.

Paaby, A. B., M. J. Blacket, A. A. Hoffmann and P. S. Schmidt, 2010 Identification of a candidate adaptive polymorphism for Drosophila life history by parallel independent clines on two continents. Mol. Ecol. **19**(4): 760–774.

Palopoli, M., M. Rockman, A. TinMaung, C. Ramsay, S. Curwen *et al.*, 2008 Molecular basis of the copulatory plug polymorphism in Caenorhabditis elegans. Nature **454**(7207): 1019–1022.

Palsson, A., A. Rouse, R. Riley-Berger, I. Dworkin and G. Gibson, 2004 Nucleotide variation in the Egfr locus of Drosophila melanogaster. Genetics **167:** 1199–1212.

Pickrell, J. K., G. Coop, J. Novembre, S. Kudaravalli, J. Z. Li *et al.*, 2009 Signals of recent positive selection in a worldwide sample of human populations. Genome Res. **19**(5): 826–837.

Rako, L., A. R. Anderson, C. M. Sgrò, A. J. Stocker and A. A. Hoffmann, 2006 The association between inversion In(3-R)Payne and clinally varying traits in Drosophila melanogaster. Genetica **128**(1–3): 373–384.

Rako, L., M. J. Blacket, S. W. McKechnie and A. A. Hoffmann, 2007 Candidate genes and thermal phenotypes: identifying ecologically important genetic variation for thermotolerance in the Australian Drosophila melanogaster cline. Mol. Ecol. **16**(14): 2948–2957.

Sackton, T. B., R. J. Kulathinal, C. M. Bergman, A. R. Quinlan, E. B. Dopman *et al.*, 2009 Population genomic inferences from sparse high-throughput sequencing of two populations of Drosophila melanogaster. Genome Biol. Evol. **1:** 449–465.

Sandrelli, F., E. Tauber, M. Pegoraro, G. Mazzotta, P. Cisotto *et al.*, 2007 A molecular basis for natural selection at the timeless locus in Drosophila melanogaster. Science **316**(5833): 1898–1900.

Schmidt, J. M., R. T. Good, B. Appleton, J. Sherrard, G. C. Raymant *et al.*, 2010 Copy number variation and transposable elements feature in recent, ongoing adaptation at the Cyp6g1 locus. PLoS Genet. **6**(6): e1000998.

Schmidt, P., C. Zhu, J. Das, M. Batavia, L. Yang *et al.*, 2008 An amino acid polymorphism in the couch potato gene forms the basis for climatic adaptation in Drosophila melanogaster. Proc. Natl. Acad. Sci. USA **105**(42): 16207.

Schmidt, P. S., D. D. Duvernell and W. F. Eanes, 2000 Adaptive evolution of a candidate gene for aging in Drosophila. Proc. Natl. Acad. Sci. USA **97**(20): 10861–10865.

Sezgin, E., D. D. Duvernell, L. M. Matzkin, Y. Duan, C. T. Zhu *et al.*, 2004 Single-locus latitudinal clines and their relationship to temperate adaptation in metabolic genes and derived alleles in *Drosophila melanogaster*. Genetics **168:** 923–931.

Sheikh, I. A., A. K. Singh, N. Singh, M. Sinha, S. B. Singh *et al.*, 2009 Structural evidence of substrate specificity in mammalian peroxidases: structure of the thiocyanate complex with lactoperoxidase and its interactions at 2.4 A resolution. J. Biol. Chem. **284**(22): 14849–14856.

Singh, N. D., P. F. Arndt and D. A. Petrov, 2005 Genomic heterogeneity of background substitutional patterns in *Drosophila melanogaster*. Genetics **169:** 709–722.

Singh, R., and A. Long, 1992 Geographic-variation in Drosophila—from molecules to morphology and back. Trends Ecol. Evol. **7**(10): 340–345.

Singh, R. S., 1989 Population genetics and evolution of species related to Drosophila melanogaster. Annu. Rev. Genet. **23:** 425–453.

Singh, R. S., and L. R. Rhomberg, 1987 A comprehensive study of genic variation in natural populations of *Drosophila melanogaster*. II. Estimates of heterozygosity and patterns of geographic differentiation. Genetics **117:** 255–271.

Slatkin, M., 1981 Estimating levels of gene flow in natural populations. Genetics **99:** 323–335.

Stark, A., J. Brennecke, N. Bushati, R. B. Russell and S. M. Cohen, 2005 Animal MicroRNAs confer robustness to gene expression and have a significant impact on 3′UTR evolution. Cell **123**(6): 1133–1146.

Storey, J., 2002 A direct approach to false discovery rates. J. R. Stat. Soc. Ser. B Stat. Methodol. **64:** 479–498.

Tajima, F., 1983 Evolutionary relationship of DNA sequences in finite populations. Genetics **105:** 437–460.

Tauber, E., M. Zordan, F. Sandrelli, M. Pegoraro, N. Osterwalder *et al.*, 2007 Natural selection favors a newly derived timeless allele in Drosophila melanogaster. Science **316**(5833): 1895–1898.

Teshima, K. M., G. Coop and M. Przeworski, 2006 How reliable are empirical genomic scans for selective sweeps? Genome Res. **16**(6): 702–712.

Tsouloufis, T., A. Mamalaki, M. Remoundos and S. J. Tzartos, 2000 Reconstitution of conformationally dependent epitopes on the N-terminal extracellular domain of the human muscle acetylcholine receptor alpha subunit expressed in Escherichia coli: implications for myasthenia gravis therapeutic approaches. Int. Immunol. **12**(9): 1255–1265.

Turner, T. L., M. T. Levine, M. L. Eckert and D. J. Begun, 2008 Genomic analysis of adaptive differentiation in *Drosophila melanogaster*. Genetics **179:** 455–473.

Tweedie, S., M. Ashburner, K. Falls, P. Leyland, P. McQuilton *et al.*, 2009 FlyBase: enhancing Drosophila Gene Ontology annotations. Nucleic Acids Res. **37**(Database issue): 555–559.

Umina, P. A., A. R. Weeks, M. R. Kearney, S. W. McKechnie and A. A. Hoffmann, 2005 A rapid shift in a classic clinal pattern in Drosophila reflecting climate change. Science **308**(5722): 691–693.

Umina, P. A., A. A. Hoffmann, A. R. Weeks and S. W. McKechnie, 2006 An independent non-linear latitudinal cline for the sn-glycerol-3-phosphate (alpha-Gpdh) polymorphism of Drosophila melanogaster from eastern Australia. Genet. Res. **87**(1): 13–21.

Voelker, R. A., C. C. Cockerham, F. M. Johnson, H. E. Schaffer, T. Mukai *et al.*, 1978 Inversions fail to account for allozyme clines. Genetics **88:** 515–527.

Voight, B. F., S. Kudaravalli, X. Wen and J. K. Pritchard, 2006 A map of recent positive selection in the human genome. PLoS Biol. **4**(3): e72.

Wang, X., D. S. Green, S. P. Roberts and J. S. de Belle, 2007 Thermal disruption of mushroom body development and odor learning in Drosophila. PLoS One **2**(11): e1125.

Wright, S., 1931 Evolution in Mendelian populations. Genetics **16:** 97–159.

Yang, Z., 1996 Among-site rate variation and its impact on phylogenetic analyses. Trends Ecol. Evol. **11**(9): 367–372.

Yang, Z., 2007 PAML 4: phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. **24**(8): 1586–1591.

Yang, Z., S. Kumar and M. Nei, 1995 A new method of inference of ancestral nucleotide and amino acid sequences. Genetics **141:** 1641–1650.

Yu, J., S. Pacifico, G. Liu and R. L. Finley, 2008 DroID: the Drosophila Interactions Database, a comprehensive resource for annotated gene and protein interactions. BMC Genomics **9**: 461.

Communicating editor: M. W. Nachman

GENETICS

Supporting Information
http://www.genetics.org/cgi/content/full/genetics.110.123059/DC1

Genomic Differentiation Between Temperate and Tropical Australian Populations of *Drosophila melanogaster*

Bryan Kolaczkowski, Andrew D. Kern, Alisha K. Holloway and David J. Begun

DOI: 10.1534/genetics.110.123059

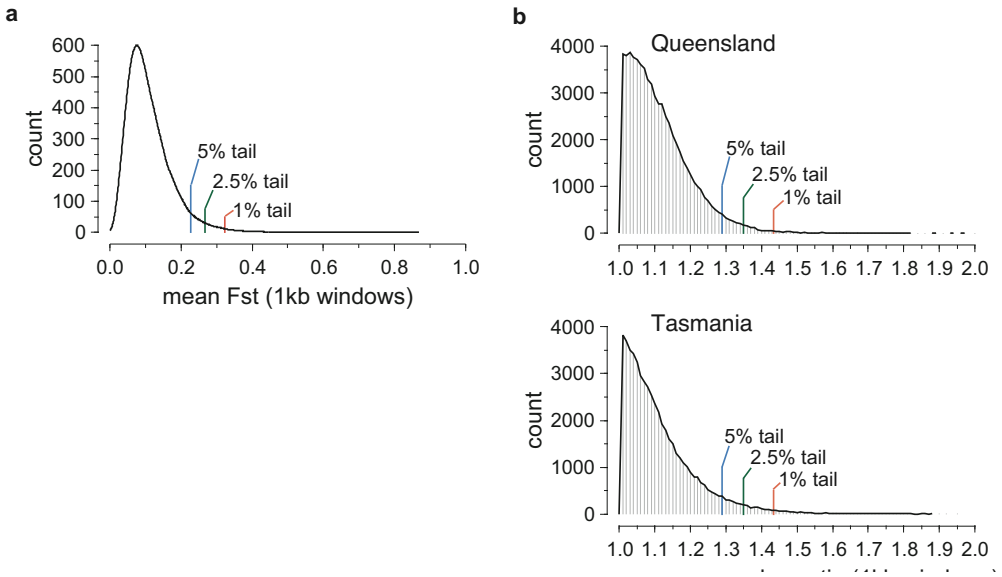

FIGURE S1.—Upper-tail F_{ST} and copy-number ratio cutoffs used in this study. We bin nonoverlapping 1kb genomic windows of F_{ST} (a) and copy-number ratio (b) and plot the number of windows in each bin. Tail cutoffs of 1, 2.5 and 5% are indicated in each panel.

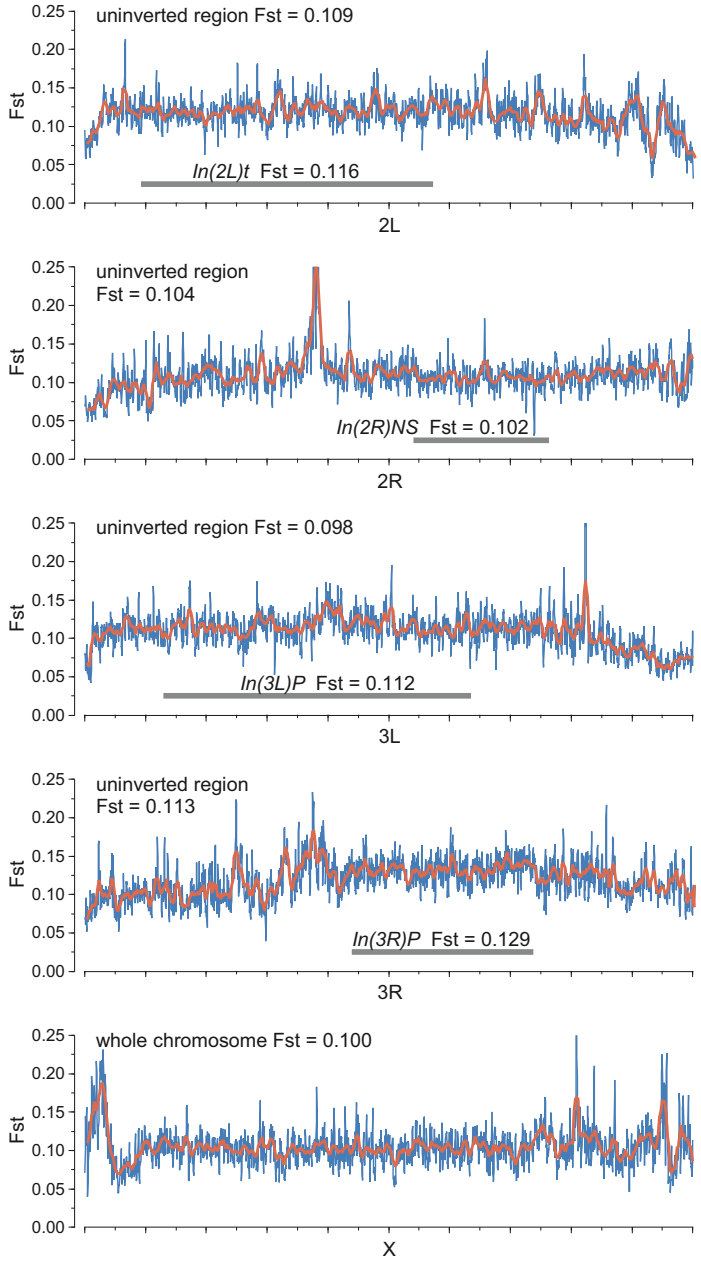

FIGURE S2.—We plot F_{ST} for standard and inverted regions of each chromosome arm. Inverted regions are indicated by gray horizontal lines. Blue series indicate average F_{ST} values over 25kb windows slid every 10kb; red lines show 200kb windows slid 50kb at a time. Overall F_{ST} across each region (standardvs. inverted) is also indicated in each panel. Note that there is no inversion on the X chromosome.

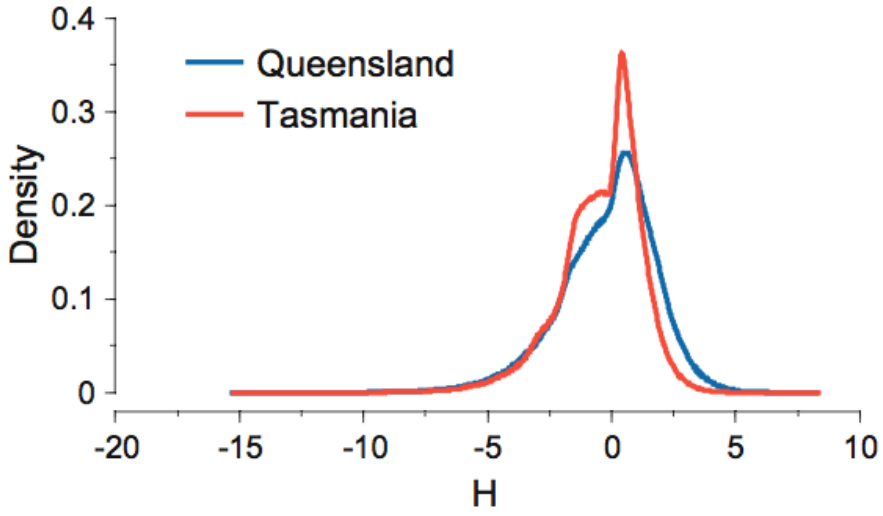

FIGURE S3.—A comparison of the distributions of H statistics computed in 1kb windows in both the Queensland and the Tasmanian samples. This is a modified version of Fay and Wu's H which exludes singleton variants. See main text for details. The medians of these distributions are significantly different from one another in a Wilcoxon rank sum test ($p < 2.2 \times 10^{-16}$).

FIGURE S4.—Fixed differences between Queensland and Tasmania at the *Irc* gene radically alter pre-mRNA structure. Computationally-inferred secondary structure of *Irc* pre-mRNA is shown for Queensland and Tasmania alleles.

TABLE S1

We report average polymorphism per 1kb (π) and Fay and Wu's *H* for each population, as well as mean F_{ST} and size of highly-differentiated genomic regions for the normally-recombining portion of each chromosome arm.

	π		Fay and Wu's H			differentiated
chromosome	Queensland	Tasmania	Queensland	Tasmania	F_{ST}	region size
2L	5.060	3.146	-0.108	-0.500	0.116	2206
2R	4.905	3.284	-0.272	-0.428	0.107	2507
3L	5.051	3.130	-0.125	-0.426	0.111	2013
3R	5.010	3.117	0.043	-0.442	0.124	3295
X	2.926	2.121	-0.500	-0.456	0.097	1899

TABLE S2

Genes in the top 2.5% of 3'UTR differentiation.

Gene ID	3'UTR F_{ST}	Gene Name	Gene ID	3'UTR F_{ST}	Gene Name
FBgn0038783	0.692282	CG4367	FBgn0035028	0.341374	Start1
FBgn0038827	0.66719	Fancd2	FBgn0259680	0.337828	Pkcdelta
FBgn0033809	0.620316	CG4630	FBgn0031681	0.337809	pgant5
FBgn0033808	0.620316	CG4627	FBgn0031861	0.335605	CG17375
FBgn0038225	0.596327	soti	FBgn0051320	0.334101	CG31320
FBgn0031491	0.561771	alpha4GT1	FBgn0038223	0.334016	CG8538
FBgn0038652	0.544105	CG7720	FBgn0259984	0.333621	kuz
FBgn0032974	0.532426	CG3651	FBgn0034820	0.331211	CG13538
FBgn0039396	0.523533	CcapR	FBgn0011289	0.327801	TfIIA-L
FBgn0038478	0.521974	cal1	FBgn0016054	0.32776	phr6-4
FBgn0038292	0.515821	CG3987	FBgn0024832	0.326232	AP-50
FBgn0032409	0.469866	Ced-12	FBgn0019968	0.32613	Khc-73
FBgn0032520	0.468198	CG10859	FBgn0031619	0.321426	CG3355
FBgn0032318	0.458057	CG14072	FBgn0085374	0.320338	CG34345
FBgn0010441	0.446403	pll	FBgn0032515	0.319843	loqs
FBgn0038124	0.436888	CG14380	FBgn0031752	0.317964	CG9044
FBgn0032465	0.430473	CG12404	FBgn0024432	0.317639	Dlc90F
FBgn0023415	0.426255	Acp32CD	FBgn0053092	0.317186	CG33092
FBgn0031284	0.424294	CG3876	FBgn0029161	0.316588	slmo
FBgn0039049	0.421011	CG6726	FBgn0085208	0.316588	CG34179
FBgn0020368	0.419943	Vha68-1	FBgn0035370	0.315363	CG1240
FBgn0043471	0.418173	kappaTry	FBgn0033128	0.315135	Tsp42Eg
FBgn0004888	0.41782	Scsalpha	FBgn0046776	0.31513	CG14033
FBgn0042710	0.414471	Hex-t2	FBgn0050050	0.314525	CG30050
FBgn0010222	0.399271	Nmdmc	FBgn0011244	0.313799	Hsp60B
FBgn0024947	0.395613	NTPase	FBgn0259966	0.311797	Sfp51E
FBgn0038819	0.393874	Cpr92F	FBgn0028482	0.3112	CG16857
FBgn0039061	0.384126	Ir	FBgn0050043	0.307662	CG30043
FBgn0050021	0.374556	skf	FBgn0033304	0.307415	Cyp6a13
FBgn0039293	0.369886	CG11851	FBgn0039969	0.306813	Fis1
FBgn0003248	0.36957	Rh2	FBgn0032474	0.302561	DnaJ-H
FBgn0026576	0.362997	CG5991	FBgn0039107	0.301822	CG10300
FBgn0014141	0.362454	cher	FBgn0040397	0.30087	CG3655
FBgn0031529	0.36213	CG9662	FBgn0020767	0.300498	Spred
FBgn0026758	0.356538	Trf2	FBgn0033806	0.300339	CG4616
FBgn0038815	0.354985	CG5466	FBgn0022073	0.299715	Thor
FBgn0086677	0.354284	jeb	FBgn0032763	0.297223	CG17568
FBgn0034263	0.353819	CG10934	_`Bgn0013433	0.296955	beat-Ia
FBgn0039238	0.352161	CG7016	FBgn0033252	0.296132	CG12769

FBgn0039239	0.352161	CG13641	FBgn0038407	0.296092	CG6126
FBgn0031522	0.344939	CG3285	FBgn0015844	0.295943	Xpd

We calculated the mean F_{ST} of each gene's 3'UTR region and list the top 2.5%.

TABLE S3

Genes in the top 2.5% of whole-gene nonsynonymous F_{ST} differentiation.

Gene ID	Nonsyn F_{ST}	Gene Name	Gene ID	Nonsyn F_{ST}	Gene Name
FBgn0031345	0.440546	CG18132	FBgn0035687	0.261204	CG13296
FBgn0086677	0.432406	jeb	FBgn0051380	0.260273	CG31380
FBgn0034288	0.41522	CG5084	FBgn0035216	0.257985	CG9168
FBgn0032331	0.401399	CG14913	FBgn0031321	0.255153	Tgt
FBgn0033093	0.38458	CG3270	FBgn0036951	0.255088	CG7017
FBgn0033697	0.384211	Cyp6t3	FBgn0039413	0.254965	CG14556
FBgn0053339	0.381734	CG33339	FBgn0025676	0.254259	CkIIalpha-i3
FBgn0053205	0.376138	CG33205	FBgn0260011	0.254035	nimC4
FBgn0035276	0.37592	CG12022	FBgn0021764	0.253939	sdk
FBgn0029173	0.373758	fu2	FBgn0038195	0.253804	CG3061
FBgn0051082	0.372544	CG31082	FBgn0003486	0.2534	spo
FBgn0025621	0.372522	CG16989	FBgn0036593	0.253169	CG13048
FBgn0085401	0.367414	CG34372	FBgn0051265	0.252429	CG31265
FBgn0011285	0.352576	S6kII	FBgn0035268	0.251505	CG8001
FBgn0051851	0.338734	CG31851	FBgn0031975	0.248371	Tg
FBgn0066101	0.331073	LpR1	FBgn0052104	0.248324	CG32104
FBgn0045476	0.329101	Gr64e	FBgn0039167	0.248287	CG17786
FBgn0015905	0.328591	ast	FBgn0031739	0.247517	CG14005
FBgn0035313	0.328007	CG13810	FBgn0032843	0.247437	CG10730
FBgn0085449	0.325412	CG34420	FBgn0015230	0.246036	Glut3
FBgn0003861	0.324953	trp	FBgn0034295	0.245982	CG10911
FBgn0051205	0.319649	CG31205	FBgn0052154	0.245111	CG32154
FBgn0039183	0.316728	Dis3	FBgn0033473	0.244758	CG12128
FBgn0038144	0.316311	CG8870	FBgn0028482	0.244381	CG16857
FBgn0052549	0.31354	CG32549	FBgn0033742	0.244029	CG8550
FBgn0038153	0.311496	Ir87a	FBgn0031782	0.243645	WDR79
FBgn0039528	0.310709	dsd	FBgn0039079	0.243057	Ir94g
FBgn0036541	0.306629	CG12486	FBgn0036320	0.242916	CG10943
FBgn0040045	0.305202	CG12460	FBgn0028852	0.242757	CG15262
FBgn0038139	0.304808	CG8795	FBgn0038125	0.242423	CG8141
FBgn0031195	0.304405	CG17600	FBgn0039201	0.242182	CG13617
FBgn0065035	0.30298	AlkB	FBgn0037989	0.24141	CG14741
FBgn0034513	0.296861	CG13423	FBgn0043796	0.241249	CG12219
FBgn0038550	0.296205	CG17801	FBgn0035918	0.241212	Cdc6
FBgn0036249	0.294651	CG11560	FBgn0037817	0.240972	Cyp12e1
FBgn0020272	0.29456	mst	FBgn0034422	0.240814	CG7137
FBgn0032045	0.291513	CG13087	FBgn0036503	0.240784	CG13454
FBgn0085213	0.290632	CG34184	FBgn0036952	0.239612	CG6933
FBgn0052751	0.290399	CG3275[illegible]	FBgn0039067	0.23927	wda

FBgn0038238	0.290022	CG14854	FBgn0085481	0.238273	CG34452
FBgn0032066	0.289623	CG9463	FBgn0043005	0.23733	CG10251
FBgn0053503	0.288901	Cyp12d1-d	FBgn0039227	0.237257	polybromo
FBgn0041607	0.288162	asparagine-synthetase	FBgn0045487	0.236246	Gr36a
FBgn0260466	0.287405	Indy-2	FBgn0260873	0.236107	CG42583
FBgn0026737	0.286948	CG6171	FBgn0032370	0.236036	CG12307
FBgn0051005	0.286607	CG31005	FBgn0020640	0.235496	Lcp65Ae
FBgn0053012	0.285977	CG33012	FBgn0010328	0.234114	woc
FBgn0030946	0.284422	CG6659	FBgn0085253	0.233799	CG34224
FBgn0085305	0.282863	CG34276	FBgn0030054	0.233653	Caf1-180
FBgn0030752	0.282683	CG9947	FBgn0038133	0.232298	Osi22
FBgn0038850	0.280486	CG17279	FBgn0039684	0.231514	Obp99d
FBgn0039246	0.279792	CG10845	FBgn0030033	0.230891	CG1387
FBgn0033599	0.279459	CG13223	FBgn0033443	0.230715	CG1698
FBgn0036948	0.277711	CG7298	FBgn0039343	0.230581	CG5111
FBgn0259199	0.277513	CG42303	FBgn0035970	0.230459	CG4483
FBgn0025866	0.277104	CalpB	FBgn0039467	0.230335	CG14253
FBgn0039083	0.276633	CG10177	FBgn0035771	0.229884	sec63
FBgn0032803	0.275137	CG13082	FBgn0038912	0.229665	CG6656
FBgn0033186	0.275072	CG1602	FBgn0033698	0.229018	CG8858
FBgn0040391	0.273132	CG2854	FBgn0014469	0.228657	Cyp4e2
FBgn0052107	0.271473	CG32107	FBgn0031961	0.227479	CG7102
FBgn0033187	0.27141	CG2144	FBgn0046689	0.227113	Takl1
FBgn0051251	0.268618	CG31251	FBgn0054041	0.226485	CG34041
FBgn0051461	0.267248	CG31461	FBgn0034141	0.22639	CG8311
FBgn0085341	0.265634	CG34312	FBgn0033850	0.226218	CG13331
FBgn0036193	0.265447	CG14135	FBgn0032142	0.225804	CG13120
FBgn0051099	0.265062	CG31099	FBgn0037533	0.225779	CD98hc
FBgn0036062	0.264058	CG6685	FBgn0039080	0.225681	Ir94h
FBgn0032484	0.263439	kek4	FBgn0031817	0.225339	CG9531
FBgn0013949	0.262919	Ela	FBgn0037248	0.225186	Spargel
FBgn0033289	0.262138	CG2121	FBgn0031429	0.224589	CG15393
FBgn0032602	0.26173	CG13278			

We calculated the mean nonsynonymous F_{ST} of each gene and list the top 2.5%

TABLE S4

Genes in the top 2.5% of individual-domain nonsynonymous F_{ST} differentiation.

Gene ID	Gene Name	Domain	Nonsyn F_{ST}
FBgn0033093	CG3270	DAO	0.38458
FBgn0038260	CG14855	MFS 1	0.384254
FBgn0033697	Cyp6t3	p450	0.384211
FBgn0034295	CG10911	DUF725	0.360936
FBgn0051380	CG31380	APH	0.342056
FBgn0003510	Sry-alpha	Serendipity A	0.33336
FBgn0045476	Gr64e	7tm 7	0.329101
FBgn0051205	CG31205	Trypsin	0.319649
FBgn0032602	CG13278	ASC	0.298351
FBgn0034513	CG13423	Peptidase C1 2	0.296861
FBgn0085213	CG34184	DM4 12	0.290632
FBgn0260466	Indy-2	Na sulph symp	0.287405
FBgn0030946	CG6659	Dpy19	0.284422
FBgn0030752	CG9947	CDC50	0.282683
FBgn0038850	CG17279	JHBP	0.280486
FBgn0028852	CG15262	NOT2 3 5	0.278706
FBgn0020377	Sr-CII	MAM	0.27817
FBgn0031305	Iris	DUF3610	0.27248
FBgn0033443	CG1698	SNF	0.268159
FBgn0051099	CG31099	DUF227	0.265062
FBgn0021764	sdk	fn3	0.263057
FBgn0032066	CG9463	Glyco hydro 38C	0.258052
FBgn0038005	Cyp313a5	p450	0.257784
FBgn0003486	spo	p450	0.2534
FBgn0038465	Irc	An peroxidase	0.247982
FBgn0054005	CG34005	DUF725	0.242942
FBgn0038541	TyrRII	7tm 1	0.239625
FBgn0054049	CG34049	CAP	0.237954
FBgn0045487	Gr36a	7tm 7	0.236246

We calculated the mean nonsynonymous F_{ST} of each PFam domain in each gene and took the most-differentiated domain as that gene's representative domain. We list list the top 2.5% of individual-domain differentiated genes.

TABLE S5

Significantly-enriched Gene Ontology categories in top 2.5% 1kb F_{ST} regions.

Biological Process		
GO accession	P-value	Description
GO:0007424	2.49E-09	open tracheal system development
GO:0007428	1.56E-06	primary branching, open tracheal system
GO:0007165	1.56E-06	signal transduction
GO:0007427	1.56E-06	epithelial cell migration, open tracheal system
GO:0002121	2.21E-06	inter-male aggressive behavior
GO:0007509	2.62E-06	mesoderm migration
GO:0042051	2.62E-06	compound eye photoreceptor development
GO:0006355	3.91E-06	regulation of transcription, DNA-dependent
GO:0007156	1.37E-05	homophilic cell adhesion
GO:0048477	1.78E-05	oogenesis
GO:0007435	1.78E-05	salivary gland morphogenesis
GO:0007507	1.82E-05	heart development
GO:0008543	2.13E-05	fibroblast growth factor receptor signaling pathway
GO:0035152	4.50E-05	regulation of tube architecture, open tracheal system
GO:0007614	8.48E-05	short-term memory
GO:0006816	8.48E-05	calcium ion transport
GO:0007431	1.54E-04	salivary gland development
GO:0007155	1.70E-04	cell adhesion
GO:0016044	2.50E-04	membrane organization
GO:0007293	4.51E-04	germarium-derived egg chamber formation
GO:0007411	4.51E-04	axon guidance
GO:0042048	4.51E-04	olfactory behavior
GO:0030707	4.51E-04	ovarian follicle cell development
GO:0008101	4.87E-04	decapentaplegic receptor signaling pathway
GO:0045570	4.87E-04	regulation of imaginal disc growth
GO:0035172	4.87E-04	hemocyte proliferation
GO:0007443	4.88E-04	Malpighian tubule morphogenesis
GO:0048813	4.88E-04	dendrite morphogenesis
GO:0048190	5.67E-04	wing disc dorsal/ventral pattern formation
GO:0007422	7.20E-04	peripheral nervous system development
GO:0016055	7.20E-04	Wnt receptor signaling pathway
GO:0007611	7.20E-04	learning or memory
GO:0007426	1.12E-03	tracheal outgrowth, open tracheal system
GO:0006813	1.21E-03	potassium ion transport
GO:0007517	1.32E-03	muscle organ development
GO:0048066	1.52E-03	pigmentation during development
GO:0007297	1.52E-03	ovarian follicle cell migration
GO:0048675	1.52E-03	axon extension

GO:0006811	1.67E-03	ion transport
GO:0007476	1.68E-03	imaginal disc-derived wing morphogenesis
GO:0030718	1.72E-03	germ-line stem cell maintenance
GO:0007619	2.08E-03	courtship behavior
GO:0045449	2.19E-03	regulation of transcription
GO:0007379	2.38E-03	segment specification
GO:0007417	3.00E-03	central nervous system development
GO:0007399	3.00E-03	nervous system development
GO:0030097	3.15E-03	hemopoiesis
GO:0007274	3.15E-03	neuromuscular synaptic transmission
GO:0007265	3.15E-03	Ras protein signal transduction
GO:0042078	3.15E-03	germ-line stem cell division
GO:0016199	3.46E-03	axon midline choice point recognition
GO:0007602	3.46E-03	phototransduction
GO:0048666	3.54E-03	neuron development
GO:0008355	3.91E-03	olfactory learning
GO:0008407	4.18E-03	bristle morphogenesis
GO:0016477	4.18E-03	cell migration
GO:0016339	4.96E-03	calcium-dependent cell-cell adhesion
GO:0055085	5.47E-03	transmembrane transport
GO:0008063	5.77E-03	Toll signaling pathway
GO:0008354	5.77E-03	germ cell migration
GO:0006357	6.28E-03	regulation of transcription from RNA polymerase II promoter
GO:0035147	6.60E-03	branch fusion, open tracheal system
GO:0008344	6.60E-03	adult locomotory behavior
GO:0009953	6.60E-03	dorsal/ventral pattern formation
GO:0035277	6.60E-03	spiracle morphogenesis, open tracheal system
GO:0006325	6.93E-03	chromatin organization
GO:0007494	6.93E-03	midgut development
GO:0002009	6.93E-03	morphogenesis of an epithelium
GO:0006468	8.66E-03	protein amino acid phosphorylation
GO:0019991	8.81E-03	septate junction assembly
GO:0007442	8.81E-03	hindgut morphogenesis
GO:0007291	8.81E-03	sperm individualization
GO:0007294	8.81E-03	germarium-derived oocyte fate determination
GO:0035071	9.76E-03	salivary gland cell autophagic cell death
GO:0007298	9.76E-03	border follicle cell migration
GO:0008104	1.11E-02	protein localization
GO:0000381	1.11E-02	regulation of alternative nuclear mRNA splicing, via spliceosome
GO:0017148	1.11E-02	negative regulation of translation
GO:0051225	1.11E-02	spindle assembly

GO:0001745	1.11E-02	compound eye morphogenesis
GO:0008360	1.12E-02	regulation of cell shape
GO:0007391	1.18E-02	dorsal closure
GO:0007498	1.35E-02	mesoderm development
GO:0007179	1.42E-02	transforming growth factor beta receptor signaling pathway
GO:0007350	1.42E-02	blastoderm segmentation
GO:0016481	1.45E-02	negative regulation of transcription
GO:0001700	1.49E-02	embryonic development via the syncytial blastoderm
GO:0007409	1.79E-02	axonogenesis
GO:0030162	1.80E-02	regulation of proteolysis
GO:0048749	1.95E-02	compound eye development
GO:0007018	1.95E-02	microtubule-based movement
GO:0007268	1.98E-02	synaptic transmission
GO:0007275	2.04E-02	multicellular organismal development
GO:0035023	2.20E-02	regulation of Rho protein signal transduction
GO:0007367	2.20E-02	segment polarity determination
GO:0006911	2.50E-02	phagocytosis, engulfment
GO:0048102	2.57E-02	autophagic cell death
GO:0009987	2.72E-02	cellular process
GO:0006096	2.82E-02	glycolysis
GO:0016318	3.31E-02	ommatidial rotation
GO:0008045	3.31E-02	motor axon guidance
GO:0030036	3.80E-02	actin cytoskeleton organization
GO:0046843	3.82E-02	dorsal appendage formation
GO:0045475	3.82E-02	locomotor rhythm
GO:0006342	3.82E-02	chromatin silencing
GO:0051726	3.82E-02	regulation of cell cycle
GO:0007474	3.82E-02	imaginal disc-derived wing vein specification
GO:0006350	3.82E-02	transcription
GO:0007186	4.03E-02	G-protein coupled receptor protein signaling pathway
GO:0000122	4.78E-02	negative regulation of transcription from RNA polymerase II promoter

Molecular Function

GO accession	P-value	Description
GO:0005515	1.57E-08	protein binding
GO:0003700	8.12E-07	transcription factor activity
GO:0005509	1.93E-04	calcium ion binding
GO:0004889	3.64E-04	nicotinic acetylcholine-activated cation-selective channel activity
GO:0003729	3.64E-04	mRNA binding
GO:0003702	3.64E-04	RNA polymerase II transcription factor activity
GO:0004871	3.64E-04	signal transducer activity

GO:0043565	8.02E-04	sequence-specific DNA binding
GO:0008188	1.87E-03	neuropeptide receptor activity
GO:0003704	1.89E-03	specific RNA polymerase II transcription factor activity
GO:0003777	3.08E-03	microtubule motor activity
GO:0005096	5.38E-03	GTPase activator activity
GO:0016566	5.92E-03	specific transcriptional repressor activity
GO:0016563	6.18E-03	transcription activator activity
GO:0005249	7.73E-03	voltage-gated potassium channel activity
GO:0003723	8.11E-03	RNA binding
GO:0003779	9.12E-03	actin binding
GO:0004888	1.02E-02	transmembrane receptor activity
GO:0005085	1.02E-02	guanyl-nucleotide exchange factor activity
GO:0008270	1.20E-02	zinc ion binding
GO:0003676	1.27E-02	nucleic acid binding
GO:0003730	1.64E-02	mRNA 3'-UTR binding
GO:0000166	1.68E-02	nucleotide binding
GO:0016251	2.52E-02	general RNA polymerase II transcription factor activity
GO:0016887	2.52E-02	ATPase activity
GO:0004725	2.97E-02	protein tyrosine phosphatase activity
GO:0005089	3.17E-02	Rho guanyl-nucleotide exchange factor activity
GO:0005524	3.65E-02	ATP binding
GO:0004674	3.67E-02	protein serine/threonine kinase activity
GO:0042623	3.67E-02	ATPase activity, coupled
GO:0005102	3.67E-02	receptor binding
GO:0004930	4.02E-02	G-protein coupled receptor activity
GO:0005516	4.16E-02	calmodulin binding
GO:0003713	4.98E-02	transcription coactivator activity

Reported P-values are corrected for a false-discovery rate of 0.05.

TABLE S6

Signficantly-enriched Gene Ontology categories in top 1% 1kb copy-number variable (CNV) regions.

Biological Process

GO accession	P-value	Description
GO:0006355	3.58E-10	regulation of transcription, DNA-dependent
GO:0007417	1.10E-08	central nervous system development
GO:0045449	6.87E-08	regulation of transcription
GO:0007507	3.73E-06	heart development
GO:0007155	1.98E-05	cell adhesion
GO:0008585	3.54E-05	female gonad development
GO:0006813	3.81E-05	potassium ion transport
GO:0048477	1.20E-04	oogenesis
GO:0001700	1.55E-04	embryonic development via the syncytial blastoderm
GO:0007399	5.02E-04	nervous system development
GO:0007304	5.70E-04	chorion-containing eggshell formation
GO:0007619	5.97E-04	courtship behavior
GO:0008587	7.14E-04	imaginal disc-derived wing margin morphogenesis
GO:0007411	7.35E-04	axon guidance
GO:0007186	8.06E-04	G-protein coupled receptor protein signaling pathway
GO:0007494	8.06E-04	midgut development
GO:0007480	9.94E-04	imaginal disc-derived leg morphogenesis
GO:0007165	1.03E-03	signal transduction
GO:0007476	1.03E-03	imaginal disc-derived wing morphogenesis
GO:0008360	1.03E-03	regulation of cell shape
GO:0045475	1.33E-03	locomotor rhythm
GO:0007422	1.39E-03	peripheral nervous system development
GO:0007224	1.53E-03	smoothened signaling pathway
GO:0048190	1.72E-03	wing disc dorsal/ventral pattern formation
GO:0042048	1.80E-03	olfactory behavior
GO:0002121	1.86E-03	inter-male aggressive behavior
GO:0007623	1.95E-03	circadian rhythm
GO:0008354	1.95E-03	germ cell migration
GO:0008104	2.26E-03	protein localization
GO:0007455	2.51E-03	eye-antennal disc morphogenesis
GO:0035023	2.87E-03	regulation of Rho protein signal transduction
GO:0016318	2.87E-03	ommatidial rotation
GO:0007400	2.87E-03	neuroblast fate determination
GO:0007015	2.96E-03	actin filament organization
GO:0008285	3.38E-03	negative regulation of cell proliferation
GO:0009987	5.31E-03	cellular process
GO:0006911	5.79E-03	phagocytosis, engulfment

GO:0016321	8.07E-03	female meiosis chromosome segregation
GO:0009790	8.83E-03	embryonic development
GO:0007419	8.83E-03	ventral cord development
GO:0048749	8.83E-03	compound eye development
GO:0007268	9.23E-03	synaptic transmission
GO:0007517	9.23E-03	muscle organ development
GO:0007163	9.23E-03	establishment or maintenance of cell polarity
GO:0008407	9.23E-03	bristle morphogenesis
GO:0016567	9.23E-03	protein ubiquitination
GO:0007498	1.17E-02	mesoderm development
GO:0006508	1.17E-02	proteolysis
GO:0006366	1.32E-02	transcription from RNA polymerase II promoter
GO:0016481	1.32E-02	negative regulation of transcription
GO:0007423	1.34E-02	sensory organ development
GO:0006350	1.42E-02	transcription
GO:0007420	1.42E-02	brain development
GO:0030707	1.71E-02	ovarian follicle cell development
GO:0007611	1.96E-02	learning or memory
GO:0030036	1.96E-02	actin cytoskeleton organization
GO:0006357	1.97E-02	regulation of transcription from RNA polymerase II promoter
GO:0016055	1.97E-02	Wnt receptor signaling pathway
GO:0007269	2.02E-02	neurotransmitter secretion
GO:0006470	2.03E-02	protein amino acid dephosphorylation
GO:0006468	2.05E-02	protein amino acid phosphorylation
GO:0007242	2.12E-02	intracellular signaling cascade
GO:0007219	2.20E-02	Notch signaling pathway
GO:0035071	2.24E-02	salivary gland cell autophagic cell death
GO:0007409	2.34E-02	axonogenesis
GO:0019730	2.34E-02	antimicrobial humoral response
GO:0000381	2.49E-02	regulation of alternative nuclear mRNA splicing, via spliceosome
GO:0007298	2.68E-02	border follicle cell migration
GO:0008355	3.04E-02	olfactory learning
GO:0000122	3.30E-02	negative regulation of transcription from RNA polymerase II promoter
GO:0016319	3.57E-02	mushroom body development
GO:0007017	4.20E-02	microtubule-based process
GO:0008283	4.71E-02	cell proliferation
GO:0006811	4.82E-02	ion transport

Molecular Function

GO accession	P-value	Description
GO:0003700	1.71E-13	transcription factor activity
GO:0005515	4.61E-13	protein binding
GO:0008270	2.69E-11	zinc ion binding
GO:0043565	7.85E-08	sequence-specific DNA binding
GO:0004879	3.69E-06	ligand-dependent nuclear receptor activity
GO:0003704	5.63E-06	specific RNA polymerase II transcription factor activity
GO:0005041	6.03E-06	low-density lipoprotein receptor activity
GO:0030528	8.75E-06	transcription regulator activity
GO:0003702	1.03E-04	RNA polymerase II transcription factor activity
GO:0008239	1.91E-04	dipeptidyl-peptidase activity
GO:0003779	1.91E-04	actin binding

GO:0003707	3.69E-04	steroid hormone receptor activity
GO:0000166	4.35E-04	nucleotide binding
GO:0016566	5.63E-04	specific transcriptional repressor activity
GO:0005249	5.95E-04	voltage-gated potassium channel activity
GO:0003729	6.33E-04	mRNA binding
GO:0004930	9.72E-04	G-protein coupled receptor activity
GO:0005089	3.13E-03	Rho guanyl-nucleotide exchange factor activity
GO:0005509	3.30E-03	calcium ion binding
GO:0004872	5.40E-03	receptor activity
GO:0004871	6.59E-03	signal transducer activity
GO:0005200	6.59E-03	structural constituent of cytoskeleton
GO:0004674	7.26E-03	protein serine/threonine kinase activity
GO:0004222	8.70E-03	metalloendopeptidase activity
GO:0042803	8.70E-03	protein homodimerization activity
GO:0008188	9.50E-03	neuropeptide receptor activity
GO:0046982	1.52E-02	protein heterodimerization activity
GO:0004725	2.36E-02	protein tyrosine phosphatase activity
GO:0005198	3.55E-02	structural molecule activity
GO:0004842	3.66E-02	ubiquitin-protein ligase activity
GO:0051082	3.66E-02	unfolded protein binding
GO:0016564	3.66E-02	transcription repressor activity
GO:0003676	4.08E-02	nucleic acid binding

Reported P-values are corrected for a false-discovery rate of 0.05.

TABLE S7

The highly-differentiated circadian-regulation gene *norpA* and its known genetic-interaction partners.

FB accession	Gene	Functional Description
FBgn0004625*	norpA	phospholypase C, temperature synchronization of circadian clock
FBgn0025680*	Cry	response to light, circadian clock synchronization
FBgn0003218*	RdgB	required for light response
FBgn0000121	Arr2	deactivation of rhodopsin-mediated signaling
FBgn0039774	CDase	synaptic transmission
FBgn0002524	lace	–
FBgn0002940*	ninaE	response to light
FBgn0085373*	rdgA	rhodopsin signaling pathway

Asterisks (*) indicate that the gene is in the top 2.5% of 1kb F_{ST} windows.

FILE S1

Corrections for pooled sampling

The nature of our experimental design creates additional noise that we must correct for in our population genetic estimates. In particular, the pooled DNA sequencing design of this manuscript creates a second level of binomial sampling, beyond what is associated with the "normal" population genetic survey. Throughout we assume that a sample of size n chromosomes is taken from nature and pooled for sequencing. This sequencing is performed to depth m which may be variable among loci/sites. Conditioning on a population frequency of the A_1 allele p, the probability of sampling i out of n A_1 alleles in our initial sample is simply binomial with parameters n and p. Thus the expected value of the sample frequency $E(i/n) = p$ and the second moment is $E((i/n)^2) = \frac{p(1-p)}{n} + p^2$.

We will first derive similar results for the pooled sampling design, and then move on to estimation of population genetic statistics. The probability of sampling k A_1 alleles out of m in our pooled sequences conditional upon having sampled i out of n in our initial sampling is again binomial, this time with parameters m and i/n. Thus the probability of sequencing k out of m reads of the A_1 allele conditional upon the population allele frequency is

$$Prob(X = k|m, n, p) = \sum_{i=0}^{n} \binom{m}{k} (i/n)^k (1 - i/n)^{m-k} \binom{n}{i} p^i (1-p)^{n-i}. \quad (1)$$

The expected value of the frequency of the allele in our sequenced sample, k/m, can then be easily found through the use of conditional expectations

$$\begin{aligned} E\{k/m\} &= E\{E\{k/m|i/n\}\} \\ &= \sum_{i=0}^{n} E\{k/m|i/n\} \times Prob(i) \\ &= \sum_{i=0}^{n} \frac{m(i/n)}{m} \binom{n}{i} p^i (1-p)^{n-i} = p \end{aligned} \quad (2)$$

We can find the second moment through similar means

$$\begin{aligned} E\{(k/m)^2\} &= E\{E\{(k/m)^2|i/n\}\} \\ &= \sum_{i=0}^{n} E\{(k/m)^2|i/n\} \times Prob(i) \\ &= \sum_{i=0}^{n} \left\{ \frac{(i/n)(1-i/n)}{m} + (i/n)^2 \right\} \times Prob(i) \\ &= \sum_{i=0}^{n} \frac{(i/n)(1-i/n)}{m} \times Prob(i) + \sum_{i=0}^{n} (i/n)^2 \times Prob(i) \end{aligned} \quad (3)$$

There are two terms in equation 3. This second term is immediately recognizable as the second moment that we examined above (i.e. $\frac{p(1-p)}{n} + p^2$). The first term after a bit of algebra turns into

$$\frac{1}{m}\{p(1-p) - \frac{p(1-p)}{n})\}$$

putting it all together, the expectation of the second moment, conditional upon the population allele frequency is

$$E\{(k/m)^2\} = p(1-p)/m - p(1-p)/mn + p(1-p)/n$$

With that result in hand we are now ready to write down the expectation of heterozygosity ($H = 2p(1-p)$) given our population allele frequency

$$\begin{aligned} E\{H\} &= E\{2p(1-p)\} = 2(E\{p\} - E\{p^2\}) \\ &= 2\Big(p - p(1-p)/m + p(1-p)/mn - p(1-p)/n\Big) \\ &= 2p(1-p)((n-1)/n)((m-1)/m) \end{aligned} \quad (4)$$

This leads to a simple bias correction on our estimates of heterozygosity which is $n/(n-1) \times m/(m-1)$. Figure S7 shows coalescent simulation results, where we generated samples from the standard coalescent model, and then resampled chromosomes with replacement to a coverage depth m. We then applied both the "double" correction derived here and the standard single correction. As can be seen in that figure, we do indeed have an unbiased estimator of heterozygosity if we correct for both the original size of our pooled sample and the coverage.

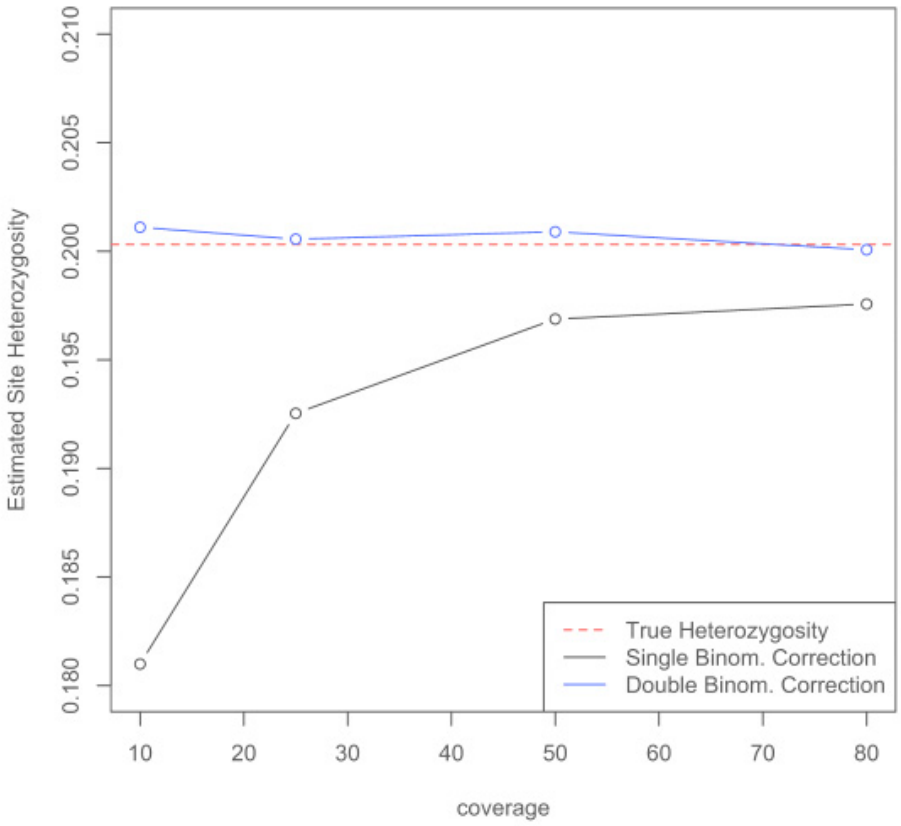

FIGURE S5.—Simulation results showing the corrected heterozygosity (eqn 4) is effective across a range of coverages used in this manuscript

Estimating θ

Of interest to us is coming up with an unbiased estimator of $\theta = 4Nu$ in the face of our pooled sampling strategy. Recently (Futschik and Schlotterer 2010) have done quite a bit of work on this problem, and they were able to come up with corrected estimators for $\theta_p i$ and θ_w. Here we derive ostensibly similar results through different means, and arrive at a generalized correction for pooled sampling which allows for construction of arbitrary estimators of θ as linear combinations of the site frequency spectrum using the system of (Achaz 2008).

We start by generalizing equation 1 of the supplement across the sample site frequency spectrum (SFS) expected under the standard neutral model. The probability of observing an allele segregating at frequency i out of n in a standard sample is $Prob(i|n) = 1/ia_n$, where $a_n = \sum_{i=1}^{n-1} 1/i$ (Ewens 2004). Thus the probability of observing an allele at frequency k out of m reads in our pooled sequence sample unconditional on the population

frequency is

$$\begin{aligned} Prob(k|m,n) &= \sum_{i=1}^{n-1} Prob(K|m,n,i)Prob(i) \\ &= \sum_{i=1}^{n-1} \binom{m}{k} (i/n)^k (1-i/n)^{m-k} (1/ia_n) \end{aligned} \tag{5}$$

This expression allows us to write down the expected number of sites segregating at frequency k out of m, call it Y_k as a function of the mutation rate θ. Conditioning on the expected number of segregating sites S in our initial sample of size n allows us to write down the expected values of the Y_ks as

$$\begin{aligned} E\{Y_k\} &= E\{S\} \times Prob(k|m,n) \\ &= \theta a_n \sum_{i=1}^{n-1} \binom{m}{k} (i/n)^k (1-i/n)^{m-k} (1/ia_n) \end{aligned} \tag{6}$$

To check the accuracy of this expression we performed coalescent simulations with a specified sample size n and mutation rate θ. The initial site frequency spectrum was recorded and then transformed to mimic our pooled sequencing by sampling alleles at each segregating site with replacement. This yields a transformed SFS Y – see figure S6.

Given the accuracy of our correction for the SFS we move on to derive a bias corrected estimation scheme for θ. In particular we note that rearrangement of equation 6 suggests a moment estimator of the type derived in Fu (1995),

$$\hat{\theta} = \frac{Y_i}{a_n} \frac{1}{Prob(k|m,n)}$$

Achaz (2009) noted that linear combinations of the SFS can be used to derive new estimators of θ given some arbitrary weighting scheme. In this context we can write down the bias corrected version of Achaz's generic estimator as

$$\hat{\theta}_\omega = \frac{1}{a_n \sum_k \omega_k} \sum_{k=1}^{m-1} \omega_k Y_k \frac{1}{Prob(k|m,n)}. \tag{7}$$

In this manuscript we focus attention on Tajima's nucleotide diversity ($\hat{\theta}_\pi$) and Fay and Wu's ($\hat{\theta}_H$) (Tajima 1983; Fay and Wu 2000). To show the potential generality of our correction scheme we present coalescent simulation results as before where we have estimated θ using six different weighting schemes: $\omega_{\pi,i}$, $\omega_{H,i}$, and $\omega_{W,i}$, each with and without use of singletons

$$\begin{aligned}\omega_{\pi,i} &= n - i \\ \omega_{H,i} &= i \\ \omega_{W,i} &= i^{-1}\end{aligned}$$

In the case where singletons are ignored each of the $\omega_1 = 0$. As can be seen in Figure S7 our bias corrected estimates are quite accurate, thus the framework we have introduced here should be general.

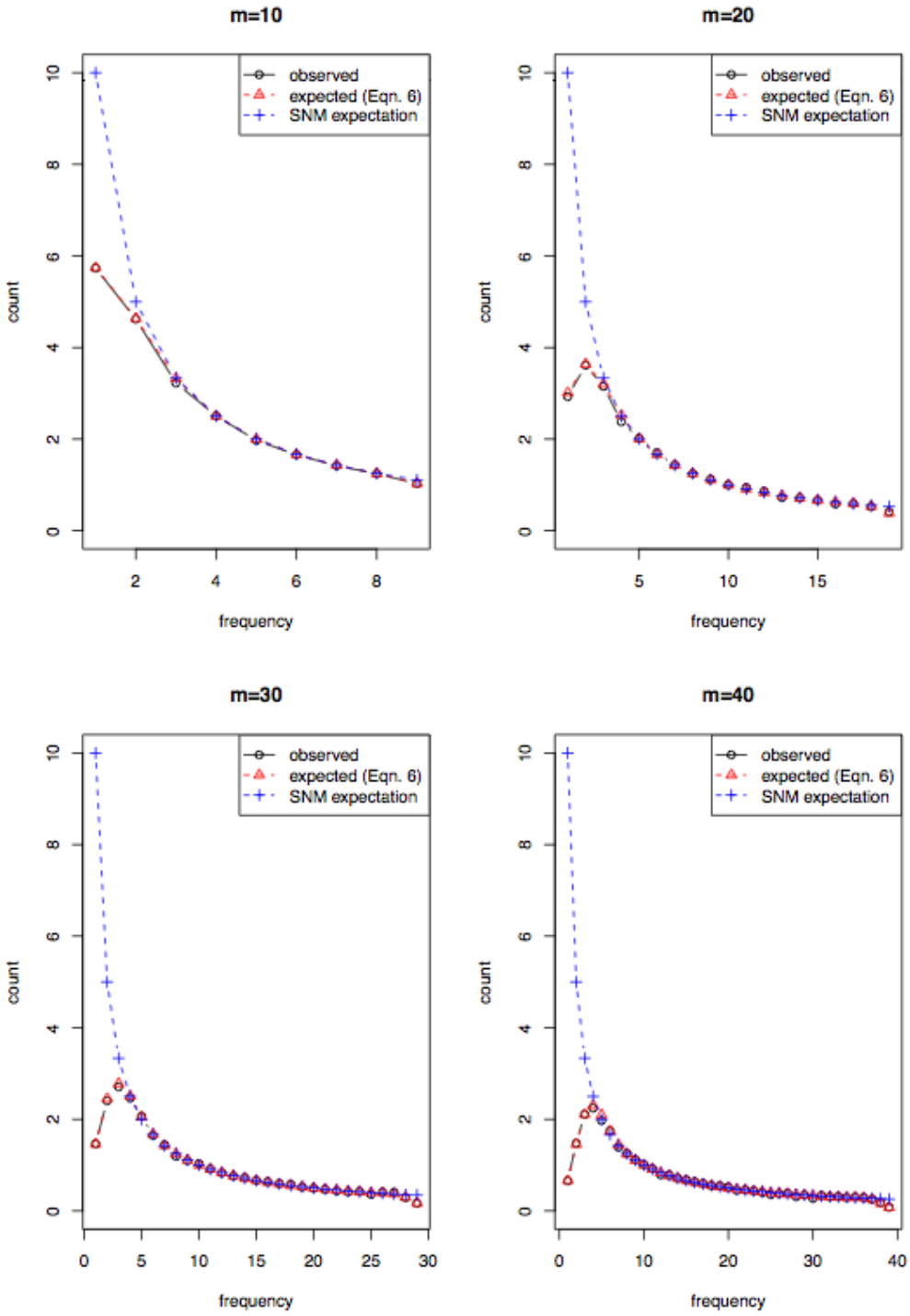

FIGURE S6.—Simulation results showing the correspondence between the observed and expected site frequency spectrum as m the sequencing depth changes. 1000 coalescent simulations were run with n=10 and θ=10. The expected values in red are derived from equation 6. Shown for comparison in blue are the expectations under the standard neutral model.

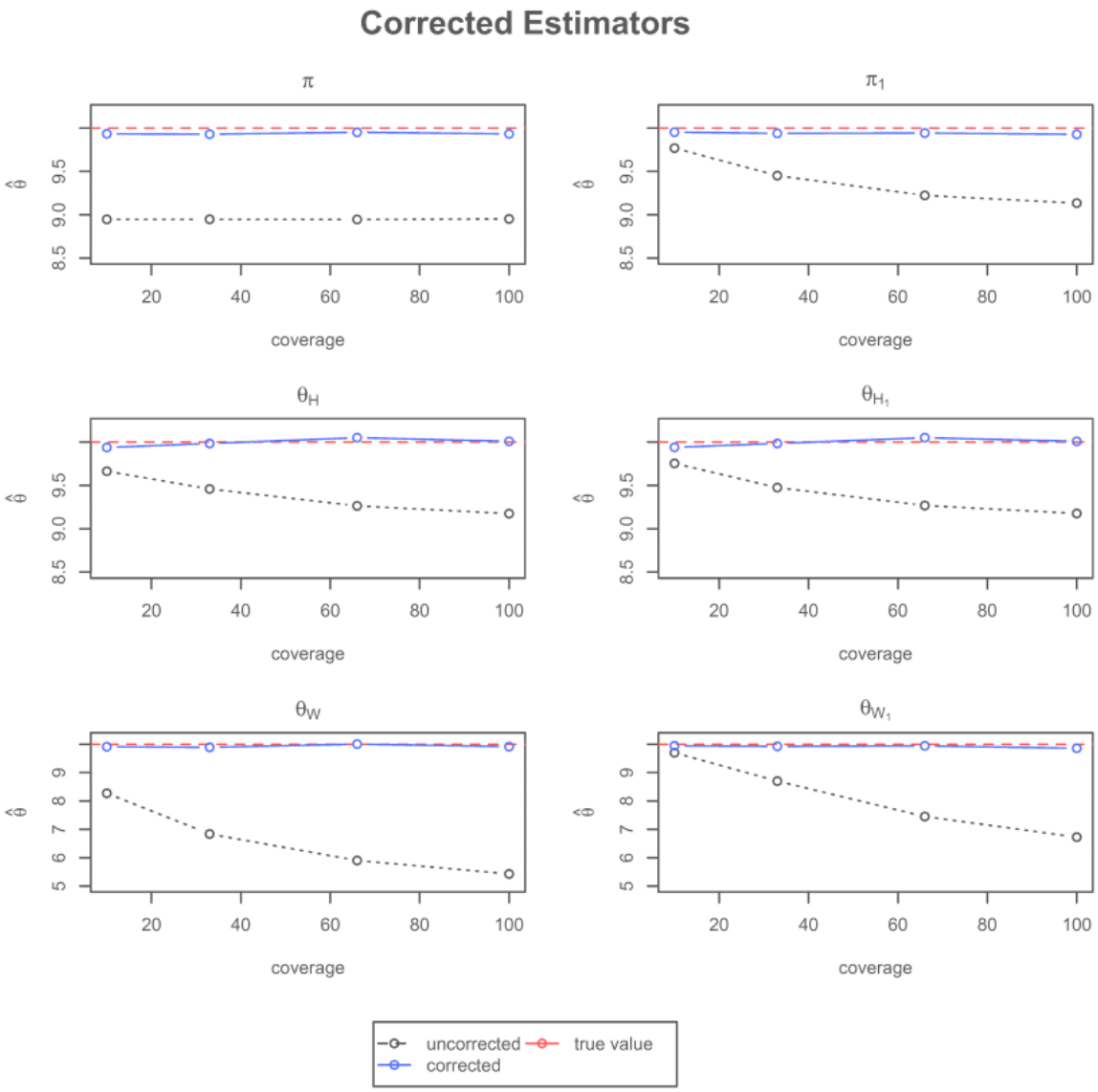

FIGURE S7.—Simulation results showing the performance of our bias corrected estimators of θ. 1000 coalescent simulations were run with $n = 40$ and $\theta = 10$. Uncorrected estimates are shown in black.

REFERENCES

ACHAZ, G., 2008, Jul) Testing for neutrality in samples with sequencing errors. Genetics *179* (3): 1409–1424.

ACHAZ, G., 2009 Frequency spectrum neutrality tests: one for all and all for one. Genetics *183* (1): 249–58.

EWENS, W. J., 2004 *Mathematical population genetics* (2nd ed ed.), Volume v. 27. New York: Springer.

FAY, J. C. and C. I. WU, 2000, Jul) Hitchhiking under positive Darwinian selection. Genetics *155* (3): 1405–13.

FU, Y. X., 1995 Statistical Properties of Segregating Sites. Theoretical Population Biology **48**: 172–197.

FUTSCHIK, A. and C. SCHLOTTERER, 2010 Massively Parallel Sequencing of Pooled DNA Samples–The Next Generation of Molecular Markers. Genetics.

TAJIMA, F., 1983, Oct) Evolutionary relationship of DNA sequences in finite populations. Genetics *105* (2): 437–60.