

Genomic Analysis of Adaptive Differentiation in *Drosophila melanogaster*

Thomas L. Turner,¹ Mia T. Levine, Melissa L. Eckert and David J. Begun

Center for Population Biology, University of California, Davis, California 95616

Manuscript received October 22, 2007

Accepted for publication March 8, 2008

ABSTRACT

Drosophila melanogaster shows clinal variation along latitudinal transects on multiple continents for several phenotypes, allozyme variants, sequence variants, and chromosome inversions. Previous investigation suggests that many such clines are due to spatially varying selection rather than demographic history, but the genomic extent of such selection is unknown. To map differentiation throughout the genome, we hybridized DNA from temperate and subtropical populations to Affymetrix tiling arrays. The dense genomic sampling of variants and low level of linkage disequilibrium in *D. melanogaster* enabled identification of many small, differentiated regions. Many regions are differentiated in parallel in the United States and Australia, strongly supporting the idea that they are influenced by spatially varying selection. Genomic differentiation is distributed nonrandomly with respect to gene function, even in regions differentiated on only one continent, providing further evidence for the role of selection. These data provide candidate genes for phenotypes known to vary clinally and implicate interesting new processes in genotype-by-environment interactions, including chorion proteins, proteins regulating meiotic recombination and segregation, gustatory and olfactory receptors, and proteins affecting synaptic function and behavior. This portrait of differentiation provides a genomic perspective on adaptation and the maintenance of variation through spatially varying selection.

THE amount and genomic distribution of polymorphism may be influenced by genetic drift, by mutation-selection balance, and by various forms of positive selection such as spatially varying selection, heterozygote advantage, or negative frequency-dependent selection. Spatially varying selection can generate allele-frequency differences between populations in spite of gene flow and lead to local adaptation, which is of particular interest as an intermediate step between intra- and interspecies variation (FELSENSTEIN 1976; ENDLER 1977; BARTON 1983).

Drosophila melanogaster has been a model system for investigating the forces maintaining polymorphism for many decades. Indeed, differentiation along latitudinal clines in this species is one of the most thoroughly documented cases of spatially varying selection (SINGH and RHOMBERG 1987; HALE and SINGH 1991). *D. melanogaster* originated in Africa and was introduced to Australia and the Americas in historic times (LACHAISE and SILVAIN 2004). Multiple phenotypes, including development time (JAMES and PARTRIDGE 1995), diapause incidence (SCHMIDT *et al.* 2005), body size (DE JONG and BOCHDANOVITS 2003), and temperature tolerance (HOFFMANN *et al.* 2002) show genetically

determined variation along latitudinal clines. Several allozyme (OAKESHOTT *et al.* 1982), DNA (GOCKEL *et al.* 2001; SEZGIN *et al.* 2004), and inversion polymorphisms (KNIBB 1982) also show clinal variation. Most putatively neutral markers show little correlation with latitude within continents, supporting the idea that clinal variation is often caused by natural selection (HALE and SINGH 1991; BERRY and KREITMAN 1993; LONG and SINGH 1995; GOCKEL *et al.* 2002; KENNINGTON *et al.* 2003). The observations that clinal variation is strongly associated with easily measurable phenotypes affecting fitness (EANES 1999; GOCKEL *et al.* 2002; CALBOLI *et al.* 2003; NORRY *et al.* 2004; KENNINGTON *et al.* 2007) and that clines for a number of phenotypes and genetic variants appear to have been independently established on multiple continents (DE JONG and BOCHDANOVITS 2003) also strongly support the proposition that many clinal variants are under spatially varying selection and that the biology of temperate and tropical populations may be quite different.

Nevertheless, these data represent a small and highly biased picture of the phenotypes and genes influenced by such selection and provide a fragmentary view of adaptive polymorphism in this genetic model system. A comprehensive genomic description of clinally differentiated genes and chromosomal regions would provide an estimate of the distribution of differentiation across the genome, which is a prerequisite for estimating the fraction of the genome influenced by local adaptation. This description would also provide new

Sequence data from this article have been deposited with the EMBL/GenBank Data Libraries under accession nos. EF670724–EF672031 and EU476191–EU476879.

¹Corresponding author: 2320 Storer Hall, 1 Shields Ave., University of California, Davis, CA 95616. E-mail: tltturner@ucdavis.edu

insights into the biological processes under selection and stimulate investigation into the functional biology of adaptive variants. Here we report estimates of differentiation with respect to latitude at ~ 3 million markers across the *D. melanogaster* genome. We used an array genotyping technique, “single feature polymorphism mapping,” which requires no SNP discovery phase (WINZELER *et al.* 1998; BOREVITZ *et al.* 2003; TURNER *et al.* 2005; WERNER *et al.* 2005; GRESHAM *et al.* 2006). This approach allowed us to map differentiation between northern and southern populations of *D. melanogaster* from the east coasts of North America and Australia, which have previously been shown to be part of latitudinal clines (SINGH and RHOMBERG 1987; HALE and SINGH 1991; KENNINGTON *et al.* 2003; ANDERSON *et al.* 2005). The resulting portrait of genomic differentiation generates candidate loci for phenotypes previously known to vary clinally and implicates interesting and unsuspected biological processes in gene-by-environment interactions.

MATERIALS AND METHODS

Fly collections: Australian flies were collected in 2004 and described in ANDERSON *et al.* (2005). Southern flies for the hybridization experiment were collected at Sorell (northern Tasmania, 41.14° S), and Miller’s Orchard (southern Tasmania, 42.46° S); northern flies were collected in northern Cardwell (Queensland), Cairns, and Cooktown (15.28° S and 16.54° S). U. S. flies were collected by Paul Schmidt in 2000. Northern flies were collected at Rocky Ridge Orchards in Bowdoinham, Maine, and Biscay Farms in Walpole, Maine. Southern flies were collected at the “Fruit and Spice Park” north of Homestead, Florida, and at the source orchards for the “Robert is Here” fruit stands south of Homestead on Route 1. Additional Australian fly lines used for sequencing (latitudes 18.16° S, 29.59° S, 32.10° S, 33.57° S, and 37.39° S) were also described in ANDERSON *et al.* (2005).

DNA isolation: We used isofemale lines from the above locations as follows: 16 lines (northern United States), 16 lines (southern United States), 17 lines (northern Australia), and 15 lines (southern Australia), pooling flash-frozen females from all lines within a region. DNA extraction of these pooled females was conducted differently for U. S. and Australian experiments. For the U. S. populations, nuclei were isolated by centrifugation, and DNA was isolated by phenol chloroform extraction followed by ethanol precipitation. For the Australian populations, DNA was extracted with QIAGEN (Valencia, CA) DNeasy kits, followed by 1:1 phenol chloroform clean up and ethanol precipitation.

Fragmentation and labeling: DNA samples (all at a volume of 30 μ l) were fragmented with a mix of DNase I (Promega, Madison, WI), One-Phor-All buffer (Amersham Biosciences), and acetylated BSA (Invitrogen, San Diego); the amounts added per sample were 4 μ l 10 \times One-Phor-All, 0.14 μ l acetylated BSA, and ~ 0.085 μ l of DNase per microgram of DNA. We created a master mix at these ratios, fragmented DNA, and then adjusted the amount of DNase in the master mix to optimize performance. Fragmentation was performed in a thermocycler at 37° for 16 min, 99° for 15 min, and 12° for 15 min [we found a large effect of different thermocyclers on the variance of fragment size, with a MJ Research PTC-200 performing much better than an Applied Biosystems (Foster City, CA) 2720]. Fragmentation performance was assessed by

running 3 μ l of fragmented DNA in a high-percentage agarose gel, with goals of a 50-bp mean fragment size, maximum intensity, and minimal variance in fragment size. When a master mix was found to perform well, all samples were fragmented with the same master mix and run in a gel together, and samples that appeared most similar were selected for labeling. Labeling was performed by adding 2 μ l of biotin-N6-ddATP (Enzo) and 3 μ l of RTdT enzyme (Promega) to each sample. RTdT was diluted from 30 to 15 units/ μ l enzyme before use by mixing RTdT enzyme, RTdT 5 \times buffer, and water at a ratio of 5:1:4. Labeling was accomplished in a thermocycler at 37° for 90 min, 99° for 15 min, and 12° for 5 min. Samples were stored at -20° until array hybridization. As with the fragmentation step, extreme care was taken to minimize variation between samples within an experiment: all samples were labeled from a single master mix at the same time.

Array hybridization: We hybridized four replicate arrays per population (16 total arrays) at the Affymetrix core facility at the University of California at Davis Genome Center. The U. S. and Australian samples were hybridized differently. For the U. S. samples, the hyb cocktail was based on Affymetrix protocols for SNP genotyping: 132 μ l 5 M tetramethylammonium chloride (TMAC); 2.2 μ l 1 M Tris, pH 7.8; 2.2 μ l 1% Tween 20; 2.2 μ l 50 mg/ml BSA; 2.2 μ l 10 mg/ml herring sperm DNA; 3.67 μ l B2 control oligo; 60 μ l labeled DNA target; and water to 220 μ l total volume. The fluidics protocol used was DNAARRAY_WS4_450. For the Australian arrays, the Affymetrix cocktail for measuring a whole-transcript double-stranded target was used: 125 μ l 2 \times hybridization mix; 17.5 μ l DMSO; 4.17 B2 control oligo; 60 μ l of labeled DNA target; water to 250 μ l. The fluidics protocol used was FS450_0001. The 2 \times hybridization mix includes the BSA and herring sperm and uses morpholineethanesulfonic acid (MES) (as in the expression arrays) instead of TMAC. We recommend the protocol used for the Australian experiment because the U. S. protocol occasionally generated chips with conspicuous spatial artifacts (which were discarded). For the Australian samples, ~ 7.5 μ g of DNA were used per microarray, while ~ 10 μ g were used in the U. S. experiment (as estimated using λ DNA gel standards and a Nanodrop spectrophotometer). All chips within each experiment were hybridized to equal amounts of DNA.

Data extraction: Arrays were scanned at the Affymetrix core facility according to the manufacturer’s specifications, which generates a binary file with intensities for each probe (a .cel file). These files were converted to text with a program available at the Affymetrix website. We used NCBI megablast to identify array probes with a single perfect match in version 4.3 of the *D. melanogaster* reference genome. Of the 6,553,600 probes on the array (which include mismatch and control oligos), 2,971,885 were retained for analysis. We used the *D. melanogaster* version 5 assembly, which contained significantly more assembled heterochromatic DNA than version 4.3, to identify probes that were unique perfect matches to the heterochromatic regions of chromosomes 2, 3, and X; this yielded an additional 32,256 probes. Thus, the total number of analyzed probes was 3,004,141.

Normalization: The goals of normalization were to control for heterogeneous average signal intensities across chips, to control for spatially nonrandom patterns of signal intensity within chips (BOREVITZ *et al.* 2003), and to remove the correlation between mean and variance of signal intensity for each probe. To spatially normalize our data, we divided each array into 1600 subarrays of 64 \times 64 probes. We log transformed raw intensity values and then divided the intensity of each oligo by the median intensity of the local 64 \times 64-probe window. This 64 \times 64-probe window was defined using all the probes on the array, so that all windows were the same size. However, medians were calculated using only the unique

probes described above. We expect most probes to hybridize equally well in all populations. Thus, the expected proportion of probes that are more intense in either population should be ~50%. After normalizing the arrays using the local 64×64 probe medians, 51.4% of probes had higher mean hybridization intensity in the southern United States. However, in Australia, 56.9% of probes were more intense in the southern United States (average difference in probe intensities were 3.1×10^{-4} in America, 3.5×10^{-3} in Australia). Therefore, we further normalized these data using quantile normalization in R (GAUTIER *et al.* 2004). After quantile normalization, 50.1 and 49.9% of probes were more intense in the southern United States and Australia, respectively, with the average difference in the means $< 4 \times 10^{-10}$ on both continents. The heterochromatic regions from the v5 annotation were added to the analysis later; arrays were renormalized to include these additional 32,356 probes, with only the data for these new probes retained.

Analysis: We estimated geographic differentiation between northern and southern populations at each probe using a two-tailed *t*-test on normalized signal intensities; all measurements were done separately for the Australian and U. S. experiments. The resulting *P*-values were converted into false discovery rates (FDR) by dividing the frequency distribution of *P*-values into 1000 bins and calculating a *q*-value for each bin (the *q*-value is the estimate of the FDR for any given *P*-value; STOREY and TIBSHIRANI 2003). Under the assumption of a uniform distribution of truly null features, the expected frequency of null features in each bin is $2,971,885$ euchromatic probes \times 0.001. Our *q*-value for each bin is simply this value divided by the number of observations in each bin (*q*-values > 1 were considered to be 1). This *q*-value is different from that of Storey and Tibshirani in that it is the expected FDR of a given range of *P*-values, rather than the FDR of all *P*-values less than a given value.

With an array probe for every 40 bp of the genome, on average, differentiated regions will often span multiple probes. To detect windows of differentiated probes, we measured the average *q*-value along each chromosome arm in windows of 20, 50, and 100 probes, moving 10% of the window size at each measurement. We generated null distributions for these overlapping windows by permuting the locations of probes on each chromosome arm 50 times. We estimated FDR for windows by dividing the expected number of windows of a given value (from the permutations) by the observed number for each window size; each chromosome arm was treated separately. Assembled heterochromatic regions are relatively small and experience little crossing over, so linked selection could lead to differentiation of entire heterochromatic regions relative to the genomic average. Significance of the fourth chromosome and heterochromatic regions of chromosomes 2, 3, and X was therefore determined by permuting all probes in the genome to generate a genomic average. The expected degree of overlap between continents was calculated under the naive assumption of independence between base pairs and is simply the percentage of the eukaryotic genome significant in Australia (1.36%) times the percentage significant in the United States (0.43%) times the number of base pairs in the eukaryotic genome (1.17×10^6).

Copy number variants (CNVs) at different frequencies in DNA pools may generate directionality in signal intensities between northern and southern samples. For example, a window corresponding to a duplicated region that is at high frequency in the north and low frequency in the south should have greater signal intensity in the north at all probes that are performing well. We conservatively considered a window to be a candidate copy number variant if 45 probes in a 50-probe window or 80 probes in a 100-probe window all had higher

means in the same population; the permutation-based 0.001 FDR threshold for this test for each chromosome arm was equal to or less than these values for all chromosomes.

We used a gene ontology (GO) database (amigo.geneontology.org/cgi-bin/amigo/go.cgi) to determine whether genes that overlap differentiated windows are nonrandomly distributed with respect to function; a gene was considered to overlap a differentiated window if any portion of its transcript was within the window. We used only terms associated with more than five genes. Significance was determined by comparing the observed proportion of genes in each GO term to an expected number, which is the binomial sampling probability of sampling an equal or larger number of genes in each category, given the number of significant genes in the genome. Because regions of the genome are significant in our analysis, and some genes with related functions are clustered in the genome, these probabilities should be interpreted with caution. To partially account for this effect, we sampled regions of the genome that exactly matched the distribution of significant regions 250 times and determined the number of genes for each GO term in these simulated data. We consider a GO term to be significant when $P < 0.05$ for both the binomial sampling test and the sampling-of-genomic-regions test.

Sequencing: To investigate how inferences made from chip data reflect underlying patterns of DNA sequence variation, we sequenced three types of regions: significantly differentiated regions, random regions, and nonsignificant regions (chosen nonrandomly). The significant regions are those that were significant in the sliding-window or single-probe analyses. The random regions were determined using a random number generator; regions were discarded if they overlapped a significant region or a TE insertion site. We also sequenced a set of regions that were not significant, but were not chosen randomly, to investigate particular aspects of the data. For example, some sequenced loci were near significant regions and others overlapped a single probe with a very low *q*-value that was not located within a significant window. DNA was isolated from each isofemale line, and primers were designed from the *D. melanogaster* genome sequence using Primer Select (Lasergene). All PCR products were cloned using TOPO TA kits (Invitrogen), and one clone from each isofemale line was sequenced in both directions. All sequences have been deposited in GenBank under accession nos. EF670724–EF672031 and EU476191–EU476879. See Table 4 for sequence lengths, sample sizes, and summary statistics at each locus. Sequence traces were assembled and edited in CodonCode (<http://www.codoncode.com>), and summary statistics were calculated in DNAsp (ROZAS *et al.* 2003). For comparisons of probe *q*-values to F_{ST} , F_{ST} was estimated for each polymorphic base that overlapped the 25-bp probe using the method of Hudson, Slatkin, and Maddison (HUDSON *et al.* 1992); probe F_{ST} values are the mean F_{ST} of all polymorphic bases within a probe. Wilcoxon tests, χ^2 tests, and Spearman rank correlations were computed in R; due to the presence of ties, rank correlation *P*-values are not exact.

For clinal analysis, we obtained sequence data from five additional Australian population samples at *Or22b*, *TjIIB*, and *dmt93B*. Five to 19 alleles per population were used; the 16.54° S and 15.28° S latitude populations were combined and considered 15.91° S for all loci, and the 32.10° S and 33.57° S latitude populations were combined and considered 32.84° S for *Or22b* because of low numbers of sequenced alleles from each population.

Quantitative real-time PCR: Genomic DNA was prepared from 10 northern Australian lines and 9 southern Australian lines. DNA concentrations were estimated and adjusted such that concentrations of all samples varied by no more than twofold. Quantitative real-time PCR (qPCR) primers were

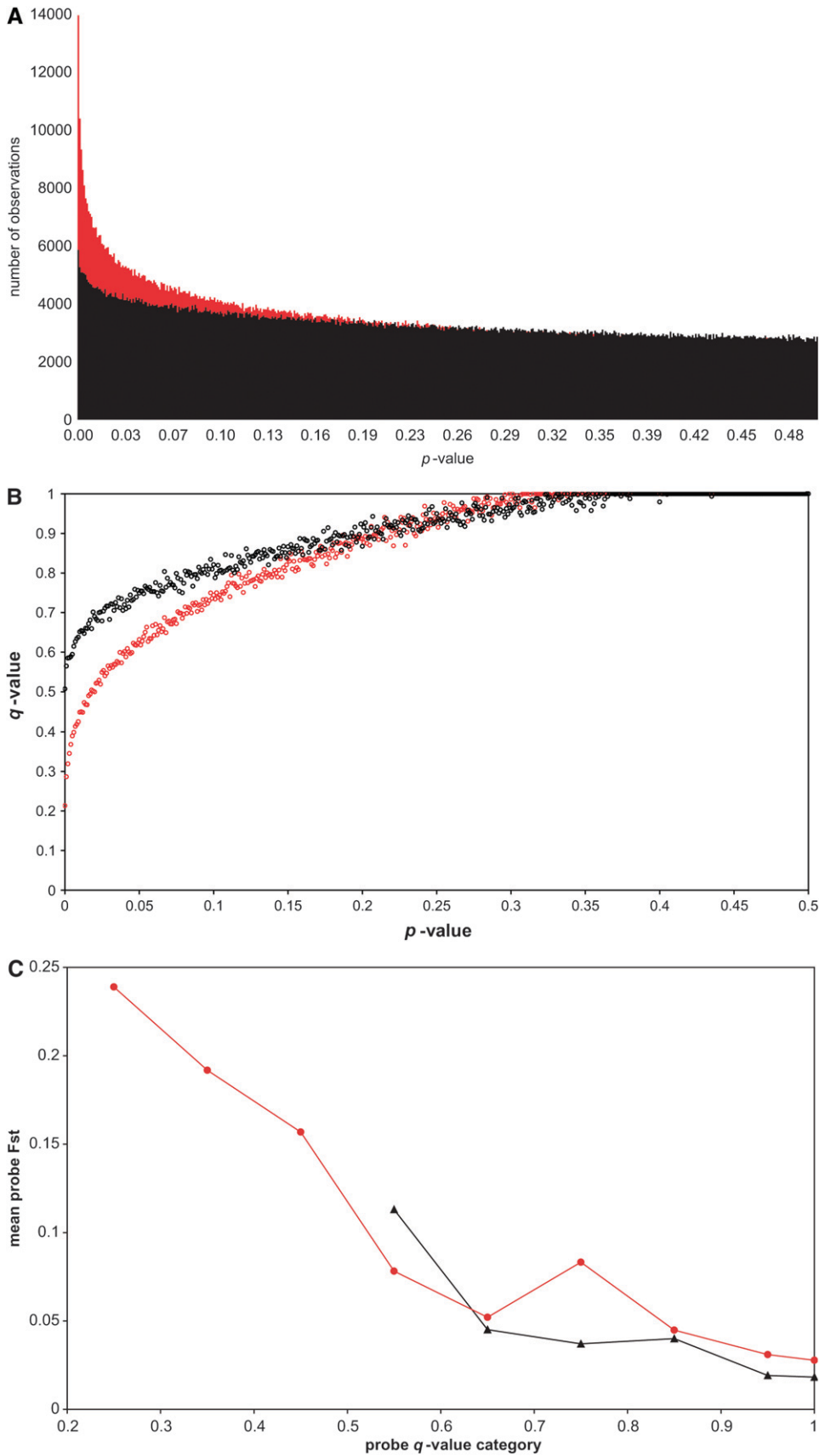


FIGURE 1.—Probe differentiation. (A) Histogram of t -test P -values. Only P -values <0.50 are shown; the remainder of the distribution is essentially flat. Red, Australian data; black, U. S. data. (B) The relationship of P -value to q -value, with Australian data in red. (C) Mean probe F_{ST} vs. probe q -value. (D) Proportion of probes with F_{ST} of zero vs. probe q -value; black triangles, U. S. data; red circles, Australian data.

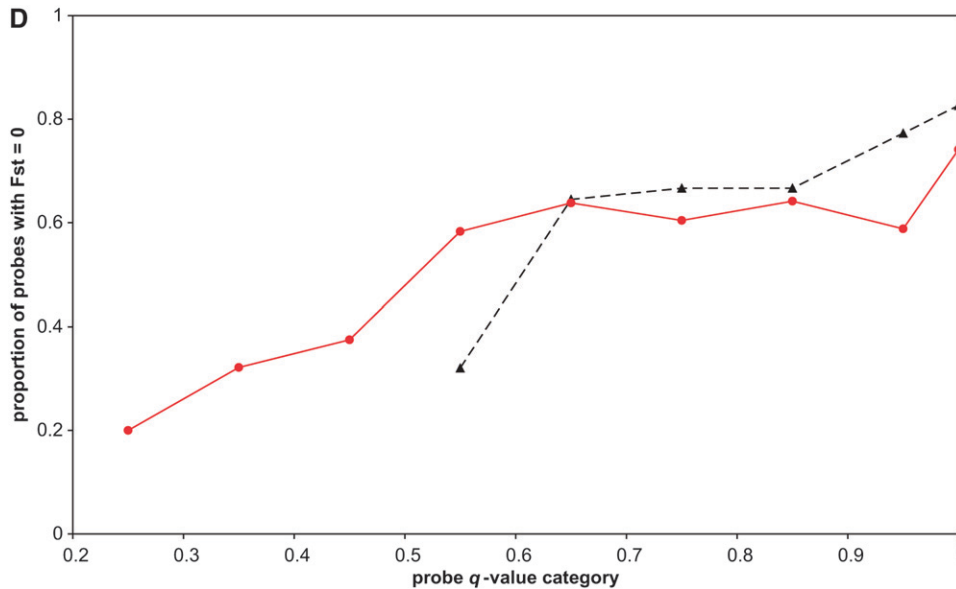


FIGURE 1.—Continued.

designed for the putative large duplication segregating on chromosome arm 3R and tested on genomic DNA; primers for two control genes, *RpL32* and *Acp53*, were taken from FIUMERA *et al.* (2005). For each DNA sample, four replicate qPCR reactions were carried out for each of the three primer pairs. The four replicates were evenly split between two plates (*i.e.*, two replicates for each DNA on each plate). We used 1 μ l of DNA plus 11 μ l of the SYBR master mix (Applied Biosystems: 6 μ l SYBR, 4 μ l dH₂O, 0.5 μ l 25 μ M forward primer, and 0.5 μ l 25 μ M reverse primer). We used both no-DNA controls and dH₂O controls. Samples were run using the default settings on an Applied Biosystems 7900HT. Data collection was performed in SDS version 2.1 and output was exported to Excel. The amount of putative copy-number DNA relative to the two control genes was determined using the $2^{-\Delta\Delta CT}$ method (LIVAK and SCHMITTGEN 2001). Significance of relative amounts of putative copy-number variant DNA between northern and southern Australian population samples was determined using the Wilcoxon test.

RESULTS AND DISCUSSION

We used Affymetrix *D. melanogaster* tiling arrays, which have a 25-bp probe for approximately every 40 bp of the genome, to identify regions showing genetic differentiation between northern and southern *D. melanogaster* populations in the United States and Australia. Because the hybridization intensity of genomic DNA to a microarray depends on sequence similarity, differentiated loci can be mapped through array hybridization (WINZELER *et al.* 1998; BOREVITZ *et al.* 2003; GRESHAM *et al.* 2006). If a polymorphism with different frequencies in two populations overlaps a probe on the array, DNA derived from these populations may hybridize to the array with different affinities (TURNER *et al.* 2005). Although this approach has several limitations, including variable sensitivity across the genome and, potentially, nonlinear change in hybridization intensity with increasing sequence mismatches (ZHANG *et al.* 2003), it provides a

useful way of measuring differentiation at a very large number of markers. We isolated DNA from pools of ≥ 15 isofemale lines from Maine (northeastern United States), Florida (southeastern United States), Queensland (northeastern Australia), and Tasmania (southeastern Australia) and hybridized each pool to four replicate arrays. We first present an analysis of differentiation at individual probes from the tiling array, followed by sliding-window analysis. Our goal here is not to formally test a population genetics model, but rather to quantify the observed distribution of differentiation across the genome. This allows us to identify the most unusual regions in the genome from populations near the endpoints of known clines and to generate hypotheses regarding genes and phenotypes harboring adaptive variation.

Differentiated probes: We used a *t*-test to measure differentiation between temperate and subtropical populations on each continent separately for the 2,971,885 array probes with single perfect matches to the *D. melanogaster* reference genome. The resulting distribution of *P*-values is shown in Figure 1A. On the basis of the excess of probes with low *P*-values, we can estimate the number of probes that overlap differentiated DNA (Figure 1A and MATERIALS AND METHODS). Assuming a uniform distribution of truly null features, 316,123 and 189,447 euchromatic probes overlap detectable differentiation in Australia and in the United States, respectively. The greater skew in the distribution of *P*-values in Australia suggests that the average differentiated probe is more differentiated between northern and southern Australian than between northern and southern U. S. populations (Figure 1A). For the majority of further analyses, we converted these *P*-values into *q*-values, which provide estimates of the false discovery rates for each range of values (see MATERIALS AND METHODS and Figure 1B).

Probes that are differentiated on both continents are excellent candidates for loci under spatially varying selection. There are 13,747 euchromatic probes in Australia with P -values < 0.001 (FDR = 21.3%) and 5838 euchromatic probes with $P < 0.001$ in the United States (FDR = 50.7%); 104 of these probes are shared ($P < 0.001$ on both continents; Figure 1A). Although the false discovery rate of these probes is high in the U. S. analysis, the combined observation of $P < 0.001$ on both continents should increase the likelihood that they are truly differentiated. The vast majority of these probes (hereafter, “golden probes”) are not located in regions of reduced crossing over and are much farther apart than the scale of linkage disequilibrium in *D. melanogaster* [(ANDOLFATTO and WALL 2003) average is 911 kb for probes on the same chromosome arm, with only one pair of probes within 500 bp of each other]. This suggests that golden probes mark many independent observations of small regions of differentiation. If differentiation at these loci was due to stochastic demographic forces or sampling variance, alternative alleles should not show parallel intensity differences with respect to climate on both continents. For example, a probe that is significantly differentiated by chance on both continents and has lower hybridization intensity in Maine should have lower hybridization in Queensland or Tasmania with equal probability. In fact, 80% of the golden probes have hybridization patterns that indicate that mismatched allele frequencies vary in a consistent manner with respect to latitude on the two continents (sign test $P < 0.001$). We sequenced 10 amplicons containing a golden probe: 8 in which hybridization intensity varied in parallel, and 2 in which we expected allele-frequency variation in opposite directions on the two continents with respect to latitude. The sequence data from these amplicons (see below) supported the array-based inferences for all 10 probes, with no apparent false discoveries.

Each of the four major autosomal arms in *D. melanogaster* segregates a paracentric inversion found throughout most of the species range. These inversions are more common in tropical populations (KNIBB 1982), with the major chromosome arm 3R inversion, *In(3R)P*, showing particularly strong clinal patterns in Australia (ANDERSON *et al.* 2005; KENNINGTON *et al.* 2006). The 104 golden probes are significantly over-represented in regions spanned by the four cosmopolitan inversions (48 observed, 30 expected, $\chi^2 = 4.15$, $P = 0.042$). This pattern is driven entirely by *In(3R)P* (26 observed, 12 expected, $\chi^2 = 5.16$, $P = 0.023$), as the golden probes are not overrepresented in inverted regions when *In(3R)P* is omitted from the analysis (22 expected, 22 observed). Thus, although some of these 104 probes may be differentiated because of linkage to *In(3R)P*, the majority are likely located near a site under spatially varying selection. A list of the locations and corresponding annotations for these 104 probes is available as supplemental data.

Correlation between array data and DNA sequence differentiation: To determine how probe q -values relate to DNA differentiation, we estimated F_{ST} between northern and southern populations for 452 25-bp probes (from 17 PCR amplicons) in the United States and for 572 probes (from 23 PCR amplicons) in Australia. Golden probe loci were not included in this analysis because they should have better-than-expected FDR due to the replicated observations of differentiation on two continents.

We found a strong, highly significant correlation between probe q -value and F_{ST} : Spearman’s rank correlations are -0.306 in Australia ($P = 6.8 \times 10^{-14}$) and -0.255 in the United States ($P = 3.8 \times 10^{-8}$). The average F_{ST} increases with lower q -value (Figure 1C), and the proportion of probes with an F_{ST} of zero decreases with lower q -value (Figure 1D). These data clearly indicate that sequence differentiation between populations is revealed by these array experiments.

The array data indicate that we should observe more differentiation in Australia than in the United States (because of the greater excess of low P -values; Figure 1A). Our sequence data show the expected pattern, as the proportion of differentiated probes and the magnitude of differentiation of the most significant probes are far greater in Australia. Because our U. S. and Australian DNA samples were hybridized to the arrays under slightly different conditions (see MATERIALS AND METHODS), comparisons between continents only on the basis of the array experiment should be interpreted cautiously. Nevertheless, the correlation between q -value and F_{ST} is similar in the two experiments, supporting the proposition that differentiation is greater between the Australian populations than between the U. S. populations. This could be due, in part, to the broader latitudinal range spanned by our Australian sample (15° S– 43° S) than by our U. S. sample (25° N– 43° N). Steeper ecological gradients in Australia, greater effective gene flow in the United States, or differences in the genetic composition of the introduced populations on the two continents could also contribute to different patterns of clinal variation on the two continents.

Sliding-window analysis: Although the probe-by-probe analysis reveals many differentiated loci, limitations of the array data motivate analyses that combine information from multiple, adjacent probes. This should dramatically increase power because the scale of linkage disequilibrium in *D. melanogaster* is often > 25 bp (HADDRILL *et al.* 2005; OMETTO *et al.* 2005). We used a sliding-window analysis to identify genomic regions associated with lower-than-expected q -values, given the distribution of q -values on each chromosome arm (SEE MATERIALS AND METHODS). Because the physical scale and magnitude of genetic differentiation is unknown, we used two significance thresholds to generate comprehensive (FDR < 0.05) and more conservative (FDR < 0.001) lists of significantly differentiated 20-, 50-, and

100-probe windows (which span ~800, 2000, and 4000 bp on average). Most significant 20- and 50-probe windows were contained within significant 100-probe windows. For example, 89% of significant DNA at $FDR < 0.001$ in Australia is significant in the 100-probe windows, with the remaining 11% composed of small, highly differentiated regions, which were not detectable with the larger window size. The union of significant overlapping windows of all sizes generated a list of nonoverlapping differentiated regions; regions composed of windows significant at $FDR < 0.05$ and $FDR < 0.001$ are hereafter referred to as regions significant at these thresholds. These thresholds merely identify outliers rather than regions having excess differentiation relative to the expectation under a population genetics model.

The window analysis confirms that a much larger proportion of the genome is significant in Australia: U. S. populations have only 31% as many base pairs in significant regions at both FDR thresholds. Although many regions are significant on only one continent (typically Australia), significant regions are shared between continents much more than would be expected by chance (150 kb are significant on both continents at $FDR < 0.001$, while only 8.7 kb are expected; $\chi^2 = 125,807$, $P < 2.2e^{-16}$). Although this expected value is probably an underestimate of the true null hypothesis (because power varies across the genome and base pairs are not independent), this analysis provides genomic support for previous conclusions that many genes vary clinally on multiple continents and are likely to be experiencing spatially varying selection (SINGH and RHOMBERG 1987; HALE and SINGH 1991).

Table 1 reports the number of base pairs in significant regions and the degree of overlap of significant regions between continents for each chromosome arm. Much of the heterogeneity between chromosome arms is attributable to a few large regions. For example, large centromere-proximal regions of the X chromosome appear to be differentiated in Australia but not in the United States. Results from the window analysis support results from the probe-level analysis in that significant regions were overrepresented on chromosome arm 3R. At the conservative FDR threshold, nearly half (47.3%) of the differentiated regions significant on both continents are on 3R (70,962 bp observed, 35,752 expected; $\chi^2 = 28,271$, $P < 2.2e^{-16}$). An example of the window analysis (50-probe windows) for 3R can be seen in Figure 2. The breakpoints of the known clinal inversion *In(3R)P* are highly significant on both continents. Overall, however, significant regions are not dramatically more abundant within the inverted region (Figure 2), consistent with previous reports suggesting that levels of linkage disequilibrium in *D. melanogaster* are not consistently high throughout *In(3R)P* (KENNINGTON *et al.* 2006). The only other inversion for which precise breakpoints are known is *In(2L)t* (ANDOLFATTO

and KREITMAN 2000). Regions near the breakpoints of this inversion are not significantly differentiated in our analysis, perhaps because the cline for *In(2L)t* is much shallower than the cline for *In(3R)P*. We detected numerous, previously known clinal loci in our window analysis, including *Est6* in Australia (COPPIN *et al.* 2007), *Hex-C* in Australia (EANES 1999), *Gdh* in the United States (EANES 1999), *hsv-omega* in Australia (ANDERSON *et al.* 2003), *UGPase* in the United States (SEZGIN *et al.* 2004), *cpo* on both continents (P. SCHMIDT, personal communication), and *period* in Australia (although this is a portion of this gene that is not known to be clinal; SAWYER *et al.* 1997) in addition to the *In(3R)P* breakpoints on both continents.

The window analysis is complementary to the single-probe analysis. While the latter may detect very short, strongly differentiated regions, the window analysis combines information from multiple probes, thereby providing more power to infer differentiation. In spite of this dual approach, inherent limitations of array performance mean that many differentiated regions will not be detected. An illustrative example is the *Adh* locus, at which clinal differentiation is limited to a single coding SNP and noncoding indel mutation (BERRY and KREITMAN 1993). Indeed, we do not detect *Adh* in our window analysis. Furthermore, we would be unlikely to detect these two mutations in our single-probe analysis because both are 1 bp away from a probe on the tiling array. Other known clinal loci were missed as well, such as *Pgm*, *Gpdh*, *Treh* (SEZGIN *et al.* 2004), and *hsp70* (BETTENCOURT *et al.* 2002). Although additional arrays could increase power, resequencing efforts will be required for complete descriptions of geographic variation in the species.

The physical extent of differentiated regions linked to sites experiencing spatially varying selection depends on several population genetics parameters, including recombination rate, age of alleles, strength of selection, and amount of gene flow. The vast majority of differentiated regions that we are able to detect with this method are several hundred to several thousand base pairs long (Table 2). For regions containing at least one gene, the mean number of genes per significant region is fewer than two, so these data offer considerable opportunity to identify genes that are targets of spatially varying selection. Additionally, a number of regions that are significant on both continents contain no genes ($n = 7$ at $FDR < 0.001$) and are candidates for selection on noncoding sequences (Table 2). However, technical factors such as probe density and probe sensitivity may also affect estimates of the size of differentiated regions, and sequence data will be required for more accurate estimates. A complete list of significantly differentiated regions is available as supplemental data.

Despite the small average size of most significant regions, several are many kilobases in length. For example, seven regions are >40 kb at the $FDR < 0.05$

TABLE 1
Number of base pairs and percentage of each chromosome arm in differentiated regions

	Australia		United States		Australia and United States	
	FDR < 0.05	FDR < 0.001	FDR < 0.05	FDR < 0.001	FDR < 0.05	FDR < 0.001
Genome ^a	7,073,732 (6.0)	1,799,732 (1.5)	2,205,237 (1.9)	565,633 (0.5)	465,977 (0.4)	150,011 (0.1)
2L	1,027,158 (4.6)	138,643 (0.6)	308,705 (1.4)	68,419 (0.3)	39,956 (0.2)	5,579 (<0.1)
2R	1,292,678 (6.2)	408,972 (2.0)	371,319 (1.8)	108,206 (0.5)	60,642 (0.3)	17,340 (0.1)
3L	1,481,758 (6.2)	301,429 (1.3)	539,349 (2.3)	152,592 (0.6)	147,495 (0.6)	41,058 (0.2)
3R	1,298,808 (4.7)	332,316 (1.2)	780,856 (2.8)	186,307 (0.7)	164,828 (0.6)	70,962 (0.3)
X	1,973,523 (8.9)	618,372 (2.8)	205,008 (0.9)	50,109 (0.2)	53,056 (0.2)	15,072 (<0.1)
Fourth	1,068,622 (83.4)	798,653 (62.3)	None	None	None	None
2Lhet	5,635 (1.5)	5,635 (1.5)	None	None	None	None
2Rhet	612,471 (18.6)	123,893 (3.8)	413,409 (12.6)	277,149 (8.4)	120,873 (3.7)	112,541 (3.4)
3Lhet	597,885 (24.4)	231,395 (9.1)	821,580 (32.1)	456,522 (17.9)	489,669 (19.2)	214,906 (8.4)
3Rhet	669,211 (26.6)	284,357 (11.3)	661,296 (26.3)	246,348 (9.8)	385,718 (15.3)	135,851 (5.4)
Xhet	4,120 (2.0)	4,120 (2.0)	None	None	None	None

Numbers in parentheses are percentages.

^a Euchromatic genome.

threshold. Two of these regions, together totaling >250 kb, are nearly contiguous and are located near the centromere of the X chromosome in Australia; another large region differentiated on both continents is near the centromere on 2R. The large size of these centromere-proximal regions may result from extensive linkage disequilibrium associated with low recombination rates. However, three differentiated regions >40 kb are not located near centromeres. One lies on chromosome 3R in Australia [outside *In(3R)P*] and contains *acetylcholine esterase (ace)*, part of *timeout*, and seven unnamed genes. Two additional large regions, each of which contains a cluster of chorion protein-coding genes, are

found in Australia. These two regions contain 9 of the 13 chorion protein-coding genes in the *D. melanogaster* genome (as well as 14 other genes), suggesting strong selection on this gene family. Although the *ace* region is also significant in the U. S. analysis, the aforementioned chorion clusters are not, again illustrating that some of the most significant regions in Australia are not significant in the United States. Large differentiated regions in areas of normal recombination may often be due to differentiated copy-number variation, which is discussed below.

Sequenced regions: To investigate how inferences made from array data reflect underlying patterns of

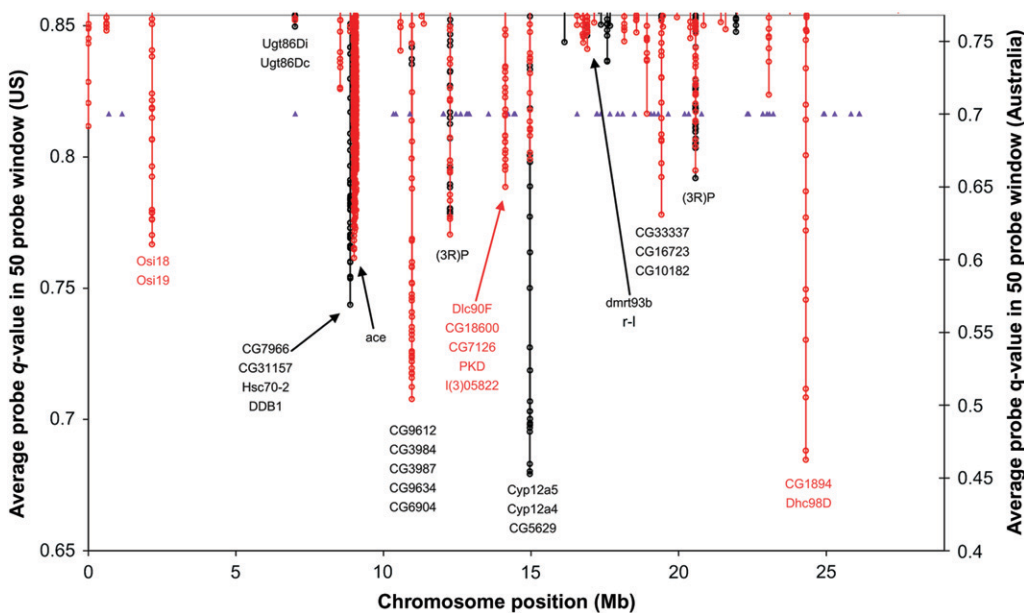


FIGURE 2.—Summary of chromosome arm 3R. Windows of 50 probes significant at FDR < 0.001 are shown in red (Australia) and black (United States). Each circle is a window; significant regions appear as vertical lines because they are small relative to the scale of a whole chromosome (with the exception of the large *ace* region). The breakpoints of *In(3R)P* are indicated, and genes that overlap some regions are shown in black type for regions significant on both continents and in red type for regions significant in only Australia. All labeled regions are also significant in the copy number variant analysis, except the *dmt93b* region. Locations of the most significant shared probes are shown as purple triangles.

TABLE 2
Summary statistics of differentiated regions

	Australia		United States		Australia and United States	
	0.05	0.001	0.05	0.001	0.05	0.001
No. of regions	1,309	252	551	115	122	25
Average size (bp)	5,404	4,919	4,002	7,142	3,819	6,000
Maximum size (bp)	140,222	31,775	110,867	108,650	20,699	19,534
Minimum size (bp)	287	292	274	312	163	682
Average gene number	1.67	1.93	1.52	1.85	1.49	1.04
No. of regions with no genes	292	58	181	36	41	7

The Australia and United States columns refer to regions significant on both continents; 0.05 and 0.001 refer to estimated false discovery rates.

DNA sequence variation, we sequenced many loci from a subset of the lines used in the array experiment. Golden probes, significantly differentiated regions from the window analysis, randomly chosen regions, and other regions that were chosen nonrandomly were all investigated. This final category includes regions that contained a probe with a low q -value on only one continent, regions located near (but not within) significant windows, and regions that were sequenced on both continents despite the fact that significant differentiation was observed on only one continent. For three regions, we extended our sampling to additional Australian populations to confirm that differentiation between cline endpoints is consistent with clinal variation.

Sequence data from 10 amplicons containing a golden probe revealed differentiation overlapping the golden probe on both continents in 9/10 cases (Table 3). In the final instance (*fas1*), the sequenced amplicon contained a single common SNP. It was differentiated in the expected direction, but was located 22 bp away from the golden probe. We suspect that this SNP may have affected hybridization of the array probe through

“proximity” effects by altering DNase fragmentation or the behavior of fragmented DNA during hybridization. Most of these loci are located in noncoding regions: differentiation may be easier to detect in these regions due to higher polymorphism, which leads to multiple differentiated SNPs or indels within the probe. An example coding region, *Gr93c*, is shown in Figure 3. A clear peak of strong differentiation located around the golden probe contains nine SNPs, three of which are nonsynonymous. These nonsynonymous SNPs are among the most differentiated sites in the gene and show parallel temperate *vs.* tropical frequency differences in the United States and Australia. Such SNPs are attractive candidates for functional investigation. Investigation of differentiated noncoding variants may provide insight into selection on gene regulation. For example, parallel differentiation at intronic sites in *cp309*, a centrosome component required for sperm motility (KAWAGUCHI and ZHENG 2004; MARTINEZ-CAMPOS *et al.* 2004); at intronic sites at *CG13466*, a predicted microtubule-related protein; and at an intergenic region near *spn-D*, which functions in female meiosis (ABDU

TABLE 3
Frequencies of sequence variants at golden probe loci

Gene	Type	Chromosome	Mb	Allele frequency				Alleles
				Tropical		Temperate		
				FL	QLD	ME	TAS	
<i>cp309</i>	Intron	3L	15.0	0.63	0.53	0.06	0.13	G → A
<i>CG13466</i>	Intron	3L	14.9	0.44	0.57	0.07	0.15	AG → TT
<i>Gr93c</i>	Coding	3R	17.7	0.63	0.75	0.25	0.25	GT → TA
<i>pointed1</i>	Intron	3R	19.2	0.73	0.71	0.38	0.23	13-bp insertion
<i>fas1</i>	Intergenic	3R	12.5	0.75	0.71	0.54	0.36	C → T (22 bp away from probe)
<i>CG32606</i>	Intergenic	X	13.8	0.25	0.18	0.06	0.00	TTA → GCG
<i>pointed2</i>	Intron	3R	19.2	0.2/0.5/0.3	0.4/0.4/0.2	0.6/0/0.4	0.8/0.2/0	GCTG → AGCG → AGCA
<i>spn-D</i>	Intergenic	3R	23.0	0.1/0/0.9	0.1/0.2/0.7	0.2/0.4/0.4	0.4/0.4/0.2	ATAA → -CGT → -TGT
<i>CG2989</i>	Coding	X	1.9	0.4/0.1/0.5	0.4/0.2/0.4	0.3/0/0.7	0.9/0/0.1	ATCC → TTCC → ACTA
<i>foraging</i>	Intron	2L	3.6	0.25	0.89	0.44	0.22	3-bp deletion

The “alleles” column shows the SNP, SNPs, or indels that overlap the probe, with three alleles present at some sites. Because these probes are our most significant, it is not surprising that most of them overlap multiple SNPs or indels. FL, Florida; QLD, Queensland; ME, Maine; TAS, Tasmania.

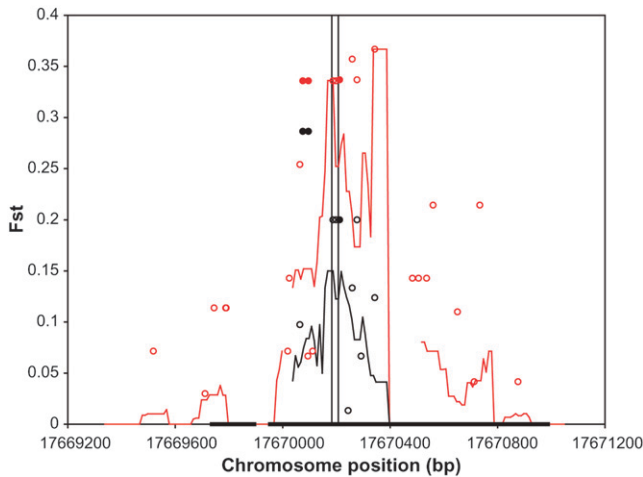


FIGURE 3.—Differentiation at *Gr93c*. Each SNP with $F_{ST} > 0$ is shown as a circle, with the three nonsynonymous SNPs as solid circles; lines are running averages of F_{ST} in 50-bp windows (Australia in red, United States in black). Thin vertical lines delineate the golden probe site. The two exons of *Gr93c* are shown as boxes.

et al. 2003), provide opportunities for investigating the functional and population genetics consequences of regulatory variation.

Table 4 shows summary statistics for all 41 amplicons sequenced from Australia and all 32 amplicons sequenced from the United States (many of these regions were also used to analyze the correlation between F_{ST} and probe q -value, but additional information is gained by analyzing these data at the scale of amplicons). The randomly chosen loci provide an unbiased sample of population genetic variation in the United States and Australia (although with only seven loci, it is limited in scope). These loci show low differentiation between northern and southern populations on each continent (average $F_{ST} = 0.054$ in Australia, 0.036 in the United States) and little differentiation between continents, with the exception of R10. This X-linked locus shows very little differentiation within continents (Table 4), but remarkably, it has 9 fixed differences between continents ($n = 19$ for each continent; $F_{ST} = 0.782$). In agreement with previous investigations (HALE and SINGH 1991), these seven random loci show no more between-continent differentiation for temperate than for tropical comparisons (average between-continent tropical *vs.* temperate $F_{ST} = 0.147$; average between-continent tropical *vs.* tropical or temperate *vs.* temperate $F_{ST} = 0.142$; Wilcoxon $P = 0.66$), which supports the hypothesis that the clines were established independently.

The conventional view of clinal variation is that it often represents the spread of a derived temperate-adapted allele to high frequency in populations farther from the equator (SEZGIN *et al.* 2004). Indeed, we observed several possible cases of this phenomenon, as indicated by much lower heterozygosity in temperate populations (*e.g.*, *Or22*, *TjIIB*, *CG13466*, and α -*tubulin*

67C). In Australia, polymorphism is lower in temperate populations in significant regions but not in random regions (Wilcoxon $P = 0.003$, 0.018 for π and θ , respectively, for significant regions). However, in many cases of strong differentiation, there is little difference in heterozygosity between temperate and tropical populations, and other significant regions have lower polymorphism in tropical populations. Further analysis regarding the histories of putatively selected alleles in such regions will be required to illuminate this issue. Finally, in Australia, the frequency spectrum is skewed toward rare alleles in temperate populations for significant but not for random regions (Wilcoxon $P = 0.0002$ for Tajima's D in significant regions). Differentiated regions in the United States show fewer systematic differences in heterozygosity or the frequency spectrum, but the trend in the U. S. data for the frequency spectrum agrees with the trend observed in Australia, with all golden probe loci having lower Tajima's D in Maine. There are individual loci in the United States that show major differences between northern and southern populations, such as the 5'-UTR of *star1*, which has a very negative skew in the frequency spectrum and a reduction in haplotype diversity in Florida, while *CG6947* shows a negatively skewed frequency spectrum and reduced polymorphism in Maine (this locus was sequenced because it contained one of the most significant probes in the genome).

As was true for the single-probe analysis, the amplicon data reveal a good correspondence between array-based and sequence-based estimates of differentiation. However, several significant regions had low F_{ST} . In some cases, this is because the differentiated region is smaller than the size of the amplicon. An example of one such narrow region of differentiation is shown in Figure 4; this region surrounds the first exon of *dmrt93B*, which is involved in sex determination (ZHU *et al.* 2000). Many of the golden probe loci have low (or zero) average F_{ST} across the sequenced amplicon, but still contain a differentiated SNP at the differentiated probe site: these loci are very attractive candidates for follow-up study, contrary to what one would be led to believe on the basis of the amplicon F_{ST} estimates in Table 4. Despite these successes, our sequencing revealed several dramatic false discoveries (such as *DmsR-2* in Australia). Although we expected < 1 in 1000 of the regions significant at $FDR < 0.001$ to be false discoveries in the technical sense, some regions may also be false discoveries in the biological sense. For example, partial homology with differentiated DNA elsewhere in the genome may affect probe intensity mapping to some loci. Two apparent false discoveries, *CG7738* and *CG9153* (Table 4), are unusual compared to the other significant sequenced regions in that they contain no probes in the lowest q -value categories. Rather, they are significantly differentiated because they possess an unusual concentration of intermediate q -values (the most significant probe over-

TABLE 4
DNA polymorphism and population differentiation in Australia and the United States

Gene	Chromosome	Location of gene (Mb)	Length of sequenced region (bp)	<i>n</i>	<i>s</i>	<i>F_{ST}</i>	Sig	Northern				Southern				<i>H_d</i> (north)	<i>H_d</i> (south)
								π	θ	π	θ	π	θ	Tajima's <i>D</i> (north)	Tajima's <i>D</i> (south)		
Australian sequence data																	
<i>Or22</i> (region 1)	2L	1.5	711	12	11	0.701	0.001	0.011	0.013	0.001	0.013	0.002	0.002	-0.057	-1.712	0.939	0.618
<i>Or22</i> (region 2)	2L	1.5	1391	19	17	0.566	0.001	0.006	0.010	0.002	0.010	0.004	0.004	-1.626	-2.097	0.994	0.895
<i>CG5470</i>	3R	11.8	744	15	15	0.390	—	0.005	0.010	0.009	0.010	0.009	0.009	-1.910	0.070	0.895	0.829
<i>Tf1B</i>	3L	4.3	1063	15	12	0.364	0.001	0.017	0.015	0.003	0.015	0.005	0.005	0.329	-1.947	0.914	0.970
<i>tefu</i>	3R	11.1	941	17	14	0.344	0.050	0.008	0.008	0.003	0.008	0.004	0.004	-0.284	-1.350	0.949	0.824
<i>CG13466</i>	3L	14.9	486	14	13	0.324	Golden	0.012	0.012	0.003	0.012	0.006	0.006	-0.031	-1.594	0.956	0.423
<i>αTub67C</i>	3L	9.7	1113	15	12	0.322	0.001	0.024	0.018	0.004	0.018	0.004	0.004	1.390	-0.100	0.990	0.939
<i>pnt</i> (region 1)	2R	19.2	462	14	12	0.287	Golden	0.008	0.005	0.007	0.005	0.006	0.006	0.515	-0.830	0.857	0.628
<i>Or59c</i>	2R	19.0	1134	13	11	0.281	0.001	0.016	0.016	0.018	0.016	0.014	0.014	-0.049	-0.707	0.987	0.982
<i>Gr93c</i>	3R	17.7	395	16	16	0.280	Golden	0.015	0.010	0.012	0.010	0.010	0.010	1.000	0.110	0.892	0.683
<i>foraging</i>	2L	3.6	474	9	15	0.251	Golden	0.023	0.022	0.012	0.022	0.018	0.018	0.217	-1.366	0.962	0.825
<i>ctm</i>	2L	7.4	1376	14	10	0.245	0.050	0.016	0.015	0.010	0.015	0.014	0.014	0.364	-1.568	0.989	1.000
<i>Gr28b</i>	2L	7.5	1310	15	13	0.221	—	0.018	0.017	0.007	0.017	0.012	0.012	0.329	-1.826	0.990	1.000
<i>CG14356</i>	3R	10.0	716	16	15	0.202	—	0.010	0.016	0.012	0.016	0.018	0.018	-1.524	-1.413	0.983	0.905
<i>Mdr49</i>	2R	8.5	841	15	13	0.180	—	0.009	0.011	0.012	0.011	0.014	0.014	-1.060	-0.652	0.962	1.000
<i>CG2989</i>	X	1.9	459	16	14	0.175	Golden	0.018	0.012	0.013	0.012	0.015	0.015	1.344	-0.946	0.900	0.923
<i>CG34008</i>	3R	16.4	1328	15	8	0.174	0.050	0.013	0.013	0.009	0.012	0.010	0.010	0.392	-0.592	0.962	1.000
<i>FasI</i>	3R	12.5	459	14	10	0.169	Golden	0.001	0.001	0.002	0.001	0.002	0.002	0.842	-0.657	0.440	0.711
<i>cher</i>	3R	12.9	975	16	15	0.168	0.001	0.009	0.011	0.010	0.011	0.013	0.013	-0.625	-1.023	0.992	1.000
<i>cp309</i>	3L	15.0	459	17	15	0.165	Golden	0.006	0.004	0.007	0.004	0.008	0.008	-0.276	-1.966	0.860	0.476
<i>spn-D</i>	3R	23.0	423	15	13	0.164	Golden	0.013	0.014	0.018	0.014	0.016	0.016	-0.316	0.675	0.838	0.923
<i>CG7747</i>	2R	11.8	1205	17	15	0.156	0.001	0.007	0.009	0.013	0.009	0.012	0.012	-0.822	0.358	0.956	0.990
R2	3L	3.0	1000	11	13	0.140	R	0.011	0.013	0.008	0.013	0.010	0.010	-0.653	-0.965	0.982	0.987
<i>dmmr193B</i>	3R	16.9	1237	10	11	0.125	0.001	0.034	0.023	0.023	0.033	0.025	0.025	0.158	-0.340	1.000	1.000
R4	3R	14.8	900	16	13	0.122	R	0.021	0.015	0.015	0.015	0.017	0.017	1.726	-0.645	0.942	0.949
<i>pnt</i> (region 2)	3R	19.2	432	17	6	0.121	Golden	0.014	0.007	0.007	0.013	0.008	0.008	0.368	-0.735	0.904	0.6
<i>intergenic-6</i>	2R	11.2	673	11	14	0.116	—	0.009	0.012	0.012	0.012	0.011	0.011	-1.288	0.279	0.945	0.978
<i>CG30457</i>	2R	12.7	930	10	13	0.091	0.001	0.019	0.013	0.013	0.017	0.012	0.012	0.550	0.203	0.978	0.795
<i>CG32606</i>	X	13.8	298	17	15	0.090	Golden	0.010	0.001	0.001	0.012	0.003	0.003	-0.736	-1.670	0.507	0.371
<i>mod1(mdg4)</i>	3R	17.2	1186	17	14	0.090	—	0.015	0.016	0.016	0.013	0.018	0.018	0.599	-0.418	0.968	0.923
R5	2R	8.3	965	15	15	0.063	R	0.007	0.009	0.009	0.010	0.009	0.009	-1.240	0.007	0.990	0.971
<i>grapes</i>	2L	16.7	1168	12	10	0.045	—	0.015	0.017	0.017	0.014	0.018	0.018	0.265	-0.470	1.000	1.000
R1	2R	10.7	1003	12	11	0.034	R	0.010	0.011	0.011	0.011	0.010	0.010	-0.057	0.476	0.985	0.982
<i>DmsR-2</i>	3L	2.3	1130	14	14	0.025	0.001	0.015	0.006	0.006	0.023	0.010	0.010	-1.594	-1.968	1.000	1.000
R6	2L	16.7	993	13	12	0.015	R	0.006	0.007	0.007	0.007	0.006	0.006	-0.490	0.741	0.910	0.924
<i>CG1942</i>	2R	3.2	1144	10	14	0.013	—	0.007	0.009	0.009	0.009	0.010	0.010	-0.902	-0.462	0.978	0.989
<i>CG9153</i>	3L	1.3	1369	14	13	0.012	0.001	0.008	0.010	0.010	0.009	0.011	0.011	-0.408	-0.554	0.989	1.000
R7	3R	21.5	980	15	13	0.006	R	0.012	0.013	0.013	0.013	0.014	0.014	-0.219	-0.138	0.971	0.962

(continued)

TABLE 4
(Continued)

Gene	Chromosome	Location of gene (Mb)	Length of sequenced region (bp)	<i>n</i>	<i>s</i>	<i>F_{ST}</i>	Sig	Northern		Southern		Tajima's <i>D</i>		<i>H_d</i>		
								π	θ	π	θ	(north)	(south)	(north)	(south)	
<i>mei-41</i>	X	16.2	1049	13	8	0.000	—	0.006	0.005	0.005	0.111	-0.014	0.949	0.786		
<i>dco</i>	3R	26.9	1192	14	8	0.000	—	0.017	0.012	0.012	-0.952	0.054	0.989	0.893		
R10	X	11.6	873	10	9	0.000	R	0.004	0.003	0.004	-1.187	-1.840	0.867	0.833		
Random average						0.054		0.011	0.010	0.010	-0.303	-0.338	0.950	0.944		
Significant average						0.233		0.013	0.009	0.010	0.027	-0.932	0.906	0.816		
						U. S. sequence data										
<i>CG6947</i>	3L	12.0	1236	14	14	0.328	—	0.002	0.006	0.006	-2.085	0.057	0.604	0.945		
<i>grapes</i>	2L	16.7	1168	13	15	0.257	0.050	0.019	0.017	0.018	0.316	-0.470	1.000	1.000		
<i>cp309</i>	3L	15.0	459	16	16	0.221	Golden	0.004	0.009	0.006	-1.585	-0.165	0.542	0.900		
<i>Gr93c</i>	3R	17.7	395	16	16	0.200	Golden	0.011	0.011	0.017	-0.010	0.438	0.817	0.917		
<i>CR15280</i>	2L	14.8	1069	11	13	0.183	0.001	0.013	0.012	0.012	0.385	0.191	0.982	0.987		
<i>star1</i>	3L	18.5	699	13	9	0.182	0.001	0.003	0.002	0.003	0.519	-1.728	0.679	0.222		
3L-1	3L	20.2	1292	13	15	0.140	—	0.004	0.002	0.005	-0.173	-1.767	0.949	0.895		
<i>pnt</i> (region 2)	3R	19.2	432	12	12	0.134	Golden	0.014	0.012	0.016	-0.303	-0.033	0.985	0.894		
R6	2L	16.7	993	8	14	0.128	R	0.008	0.007	0.009	-0.510	0.090	1.000	0.989		
<i>CG13466</i>	3L	14.9	486	15	16	0.120	Golden	0.004	0.012	0.007	-1.582	-0.707	0.790	0.975		
<i>Mob1</i>	3L	11.5	974	14	10	0.080	0.001	0.022	0.032	0.017	1.360	1.130	0.978	0.978		
<i>dmt93B</i>	3R	16.9	1237	15	16	0.080	0.001	0.024	0.034	0.025	-0.301	1.530	1.000	0.992		
<i>chm</i>	2L	7.4	1363	20	15	0.079	0.001	0.019	0.016	0.018	0.132	-0.293	0.989	1.000		
<i>mei-41</i>	X	16.2	1049	13	10	0.078	0.050	0.008	0.006	0.009	-0.728	-1.303	0.987	0.964		
<i>egl</i>	2R	19.3	821	16	16	0.066	0.001	0.008	0.005	0.009	-0.343	-1.830	0.933	0.950		
<i>pnt</i> (region 1)	3R	19.2	462	14	15	0.064	Golden	0.007	0.008	0.008	-0.598	-0.410	0.835	0.952		
<i>foi</i>	3L	8.5	1270	15	16	0.057	—	0.006	0.009	0.010	-1.410	-0.626	0.971	1.000		
R4	3R	14.8	900	13	15	0.054	R	0.015	0.022	0.020	-1.240	0.146	1.000	0.990		
<i>foraging</i>	2L	3.6	474	13	16	0.044	Golden	0.008	0.008	0.010	-1.220	-1.130	0.889	0.905		
<i>CG32606</i>	X	13.8	298	16	12	0.033	Golden	0.005	0.014	0.011	-2.211	-0.432	0.350	0.833		
R1	2R	10.7	1003	13	8	0.032	R	0.010	0.008	0.011	-0.509	-0.156	1.000	0.964		
R7	3R	21.5	981	11	11	0.027	R	0.009	0.012	0.009	0.123	0.380	0.945	0.982		
<i>Fs-scb</i>	2R	10.8	1361	16	16	0.025	—	0.003	0.005	0.005	-1.539	-0.102	0.908	0.967		
<i>spn-D</i>	3R	23.0	423	16	16	0.024	Golden	0.017	0.013	0.017	-0.004	0.639	0.967	0.817		
<i>CG5224</i>	2R	14.0	1307	16	14	0.021	0.050	0.019	0.020	0.015	1.070	0.240	0.983	1.000		
R5	2R	8.3	965	14	15	0.008	R	0.004	0.007	0.008	-2.186	-1.362	0.934	0.971		
<i>CG2989</i>	X	1.9	459	12	13	0.000	Golden	0.016	0.015	0.018	-0.161	1.290	0.955	0.910		
<i>FasI</i>	3R	12.5	457	13	12	0.000	Golden	0.003	0.002	0.004	-0.161	-1.291	0.718	0.636		
<i>CG7738</i>	2R	6.8	818	8	11	0.000	0.001	0.025	0.020	0.020	1.163	0.084	0.964	0.982		
<i>bol</i>	3L	9.1	1260	16	15	0.000	—	0.007	0.008	0.008	-0.313	-0.384	0.992	1.000		
R10	X	11.6	873	9	13	0.000	R	0.005	0.004	0.006	-0.892	-0.967	0.944	0.718		

(continued)

TABLE 4
(Continued)

Gene	Chromosome	Location of gene (Mb)	Length of sequenced region (bp)	<i>n</i>	<i>s</i>	F_{ST}	Sig	Northern		Southern		Tajima's <i>D</i> (north)	Tajima's <i>D</i> (south)	H_d (north)	H_d (south)
								π	θ	π	θ				
R2	3L	3.0	1000	9	15	0.000	R	0.010	0.009	0.009	0.011	-0.416	0.079	0.917	0.981
Random average						0.036		0.009	0.010	0.010	0.011	-0.804	-0.256	0.963	0.942
Significant average						0.093		0.012	0.013	0.013	0.013	-0.195	-0.213	0.867	0.891

Summary statistics are shown for each sequenced amplicon. "Sig" indicates whether this region is a significant region at the 0.05 or 0.001 thresholds or a golden probe; "R" indicates that the locus was chosen randomly. Loci that remain significant after eliminating information from probes with q -values > 0.6 are in boldface type. Loci that are not significant or random were sequenced for other reasons and should not be considered representative of the genome. *n*, number of chromosomes from the northern population; *s*, number of chromosomes from the southern population; Hd, haplotype diversity.

lapping the *CG7738* amplicon has a q -value of 0.68, but only 5 of 22 probes overlapping this amplicon have a q -value of 1.00). A possible explanation for this result is that our array experiment (which provides estimates of frequency differences based on a large pooled sample) detected many slightly differentiated SNPs at these loci, but this differentiation is too slight to be detected in our small sample of sequenced alleles.

On the basis of the observed relationship between amplicon F_{ST} estimates and q -values, we designed an additional *ad hoc* analysis, which identifies highly differentiated loci with fewer false discoveries. We made all probe q -values > 0.6 equal to 1.0, such that our window analysis would only detect clustering of the most differentiated probes. This modification significantly increased the Spearman's rank correlation between amplicon F_{ST} and average q -value in the U. S. data: considering all q -values, $\rho = -0.33$, $P = 0.17$; considering $q > 0.6 = 1.0$, $\rho = -0.59$, $P = 0.008$. The Australia correlation also increased: considering all q -values, $\rho = -0.38$, $P = 0.046$; considering $q > 0.6 = 1.0$, $\rho = -0.46$, $P = 0.014$. Thus, this may be a good approach for finding the most differentiated regions with the lowest false discovery rate. Alternatively, including all q -values may be the most powerful way to delineate all significantly differentiated regions for genomic analyses such as the gene ontology analysis presented below.

Despite the demonstrated usefulness of these data, which vastly increase the number of potentially clinal loci, we emphasize that the FDR estimates for regions are inaccurate, in contrast to the FDR estimates for individual probes, which appear to be very accurate. We therefore consider our three significance thresholds for regions (FDR < 0.05 , FDR < 0.001 , and the most conservative analysis just discussed) to provide three relative levels of confidence, which can be used by the community for exploratory analysis of clinal differentiation. An example of how these data may be used to gather evidence bearing on spatially varying selection of genes or classes of genes can be found in the accompanying article analyzing variation in chromatin-remodeling genes (LEVINE and BEGUN 2008, this issue). Results from all thresholds from the array analyses are available as supplemental data.

Clinal samples: Although the array analysis included only populations from the endpoints of known clines, we expect that most of the differentiated regions detected between these populations will prove to be clinal. The presence of known clinal loci in our data supports this assertion. Previous investigations of cosmopolitan *D. melanogaster* populations, especially in Australia, have found little other differentiation between populations [although one previous study found that Tasmanian populations were differentiated from southern mainland populations, additional studies did not confirm this result (AGIS and SCHLOTTERER 2001; KENNINGTON *et al.* 2003)]. To investigate this issue, we sequenced

TABLE 5
Summary of significant GO terms

GO term	Continent
Acetylcholine receptor activity	Australia, Common
Neuropeptide receptor activity	Australia
Synapse	Australia
Insect chorion formation	Australia
Structural constituent of chorion	Australia
Germline cell division	Australia
Germ-band shortening	Australia
Progression of morphogenetic furrow during eye morphogenesis	Australia
Microtubule-associated complex	Australia
Microtubule binding	Australia
Flight behavior	Australia
Glucose transport	Australia
Polypyrimidine tract binding	Australia
Defense response	Common
Response to toxin	Common
Terminal region determination	Common
Cell cycle checkpoint	United States
GABA receptor activity	United States
Negative regulation of cell proliferation	United States
Spermatogenesis	United States
Postsynaptic membrane	United States, Australia
Nerve–nerve synapse transmission	United States, Australia
Nicotinic acetylcholine-activated cation-selective channel activity	United States, Australia, Common
Glucuronosyltransferase activity	United States, Common
Polysaccharide metabolic process	United States, Common
Steroid metabolic process	United States, Common

When a term is significant in more than one data set, the *P*-value listed is the smallest of the data sets; *fungus defense* and *chromatin* are marginally significant. Common is the data set of regions significant on both continents.

three differentiated Australian loci (*Or22b*, *TjIIB*, and *dmrt93B*) in Australian populations from additional latitudes (see MATERIALS AND METHODS). All 11 common polymorphisms in the *Or22b* locus and 37/43 common polymorphisms in the *TjIIB* locus are significantly ($P < 0.05$) clinal, with r^2 values up to 0.996. Of the 20 polymorphisms falling within the most differentiated window of *dmrt93B* (Figure 4), 16 had significant r^2 values ($P < 0.05$). When the allele frequencies for all polymorphisms in each locus are averaged, all three loci remain significantly clinal (Figure 5). We did not detect a break between the Tasmanian populations used in the array analysis and the other southern Australian populations. These data support the notion that many of the highly differentiated regions identified in the array

analysis harbor clinal variation. The analysis presented in the accompanying article (LEVINE and BEGUN 2008) also supports the hypothesis that strongly differentiated regions are often clinal.

Differentiation of functional categories: Given that most significant windows are associated with one or two genes, functional annotations corresponding to genes in significant windows can be used to investigate whether certain GO terms are associated with differentiated regions (see MATERIALS AND METHODS). The results are summarized in Table 5, while the full list of significant GO terms is available as supplemental data. Genes are not distributed randomly along chromosomes with respect to function, so that physical clustering of related genes could inflate the observed significance of GO terms. However, all of the GO terms in Table 5 include genes from different locations in the genome, which supports the hypothesis that these functions are under selection. We also assessed significance of GO terms using two complementary strategies, one of which explicitly considered the clustering of related genes (see MATERIALS AND METHODS). In this section, we highlight several examples to illustrate the nonrandom nature of differentiation with respect to gene function.

Regions that are differentiated on both continents contain a dramatic overrepresentation of several gene families, including UDP glucotransferases and cytochrome P450s (cyps). A region on 3R containing five genes with glucuronosyltransferase activity and two unlinked related genes (*CG10170* and *CG30438*) is significant on both continents, leading to a highly significant overrepresentation of genes with glucuronosyltransferase activity, polysaccharide metabolism, and toxin response (Table 5). Cytochrome P450 (cyp) genes are overrepresented in shared significant regions as well: *Cyp12a4* and *Cyp12a5* are in one of the most significant regions of the genome (Figure 2). Five cyp genes are significant in additional shared regions: *Cyp6a17*, *Cyp6a22*, and *Cyp6a23* are in a significant cluster on 2R, *Cyp6g1* is also on 2R, and *Cyp6t1* is on the X (*Cyp6a17* and an additional gene on 2R, *Cyp12d1*, each contain 1 of the 104 golden probes). These cytochromes and transferases lead to the significant overrepresentation of genes involved in steroid metabolism and defense response (which also includes the significant genes *hsc70-2*, *cactin*, and *turandot C*).

The nonrandom distribution of gene functions in regions significant only in Australia provides evidence that these functions are under spatially varying selection on that continent. For example, GO terms associated with the chorion (the outer layer of the eggshell) are overrepresented in Australia. The six chorion proteins that are significant in the 3L and X chorion clusters in Australia represent 75% of the genes in the “structural constituent of chorion” GO term. Additionally, *fs(1)N*, which is known to function in chorion formation, is differentiated in Australia, contributing to the signifi-

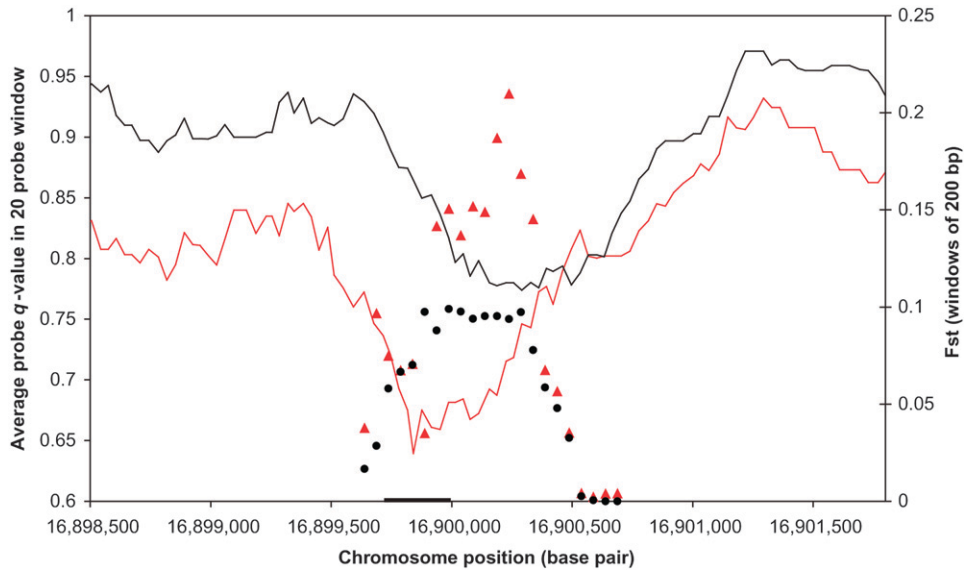


FIGURE 4.—Differentiation at *dmrt93b*. Microarray differentiation is shown for windows of 20 probes. Australia, in red, is more differentiated; the United States is shown in black. Sequence differentiation (F_{ST}) is shown as triangles for the United States and circles for Australia (each point corresponds to a 200-bp sliding window, moving 100 bp at a time). Differentiation overlaps the first exon of *dmrt93b* (box) and part of the first intron.

cant overrepresentation of the “insect chorion formation” GO category. Three other genes, *mip130*, *rbf*, and *dup*, are all involved in chorion gene amplification and are significant in Australia. Finally, three additional chorion genes, *Cp7Fa*, *Cp7Fb*, and *Cp7Fc*, are significant in the X-linked chorion cluster but are not associated with any GO terms. Interestingly, none of these 13 chorion-related genes overlaps a significantly differentiated region in the United States. Because the chorion clusters are polytenized in some tissues in females, it is possible that these regions are significant because of geographic variation in the degree of polytenization in females or different proportions of polytenized tissues in these females rather than from different allele frequencies at these genes.

Of the nine known acetylcholine receptors in the genome, portions of three (in addition to *ace* itself) are significant on both continents and an additional three are significant only in Australia. These data, and the fact that many other significant genes are involved in synaptic transmission, lead to a significant overrepresentation of a suite of synapse-related GO terms. The known examples of insecticide resistance alleles associated with acetylcholine metabolism in flies and mosquitoes (FOURNIER *et al.* 1989; WEILL *et al.* 2002, 2003) and the aforementioned significant *Cyp6g1*, which is also implicated in insecticide resistance (CHUNG *et al.* 2007; DABORN *et al.* 2007), suggest that selection pressures associated with insecticide or other environmental toxins vary with latitude.

In the United States, the cell cycle/DNA damage checkpoint genes *mei-41* and *grapes* (LAROCQUE *et al.* 2007) were both significant; the genes are unlinked, and differentiation of both is unlikely to be due to chance ($P < 0.04$ in GO analysis). The differentiated region of *mei-41* included only one nonsynonymous variant; the ancestral allele was present in all other sequenced

Drosophila genomes, which supports the idea that the derived allele is functionally important. These genes were not significant in Australia, and sequencing verified that they are differentiated only in the United States. Differentiated genes thought to play a role in double-strand break repair, recombination, or chromosome segregation in Australia include α -*tubulin 67C* (MATTHIES *et al.* 1999), *egl* (CARPENTER 1994), *tefu* (OIKEMUS *et al.* 2006), and *glu* (STEFFENSEN *et al.* 2001), with α -*tubulin 67C* and *tefu* among the most differentiated loci sequenced (*glu* sequence data are presented in LEVINE and BEGUN 2008). The verified golden probe near *spn-D* could be related to these functions as well (ABDU *et al.* 2003). In general, clinal variation could plausibly be associated with temperature variation (directly or indirectly) or with other ecological variables correlated with latitude. Alternatively, the genetics of *D. melanogaster* populations may vary with latitude as a result of clines for chromosome inversions, which are major modifiers of recombination that could result in selection on genes that regulate crossing over. In tropical populations, autosomes are frequently heterozygous for paracentric chromosome inversions, which should lead to greater frequencies of achiasmatic meioses and higher rates of nondisjunction (for autosomes). Indeed, females heterozygous for inversions on all four autosomal chromosome arms suffer a considerable drop in fertility (STALKER 1976). In the absence of data on phenotypic variation associated with these genes, several hypotheses for selection are plausible; at the present time, we can only speculate.

The wide range of overrepresented functions illustrates the pervasive effect of the environment on *D. melanogaster* biology and clearly indicates that functional analysis of natural variants can contribute to annotation of the *Drosophila* genome. Differentiation of sets of related genes or genes that interact with one another

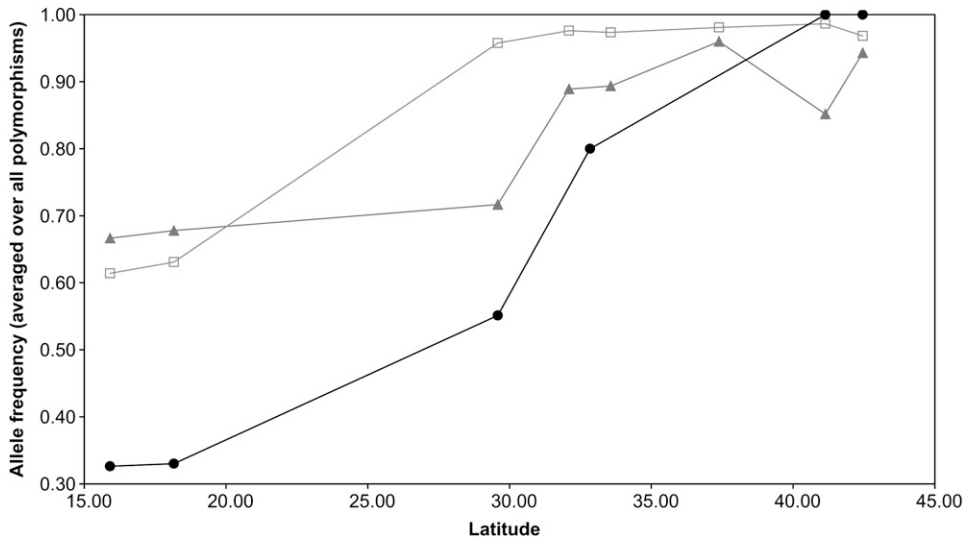


FIGURE 5.—Clinal differentiation. We sequenced *Or22b* (circles), *TjIIB* (squares), and *dmrt93B* (triangles) from additional latitudes to determine if differentiated regions were clinal. The average allele frequency for all polymorphisms in each locus is shown; multiple polymorphisms in complete linkage were considered once when taking the average. The r^2 values for these averages *vs.* latitude are 0.982, 0.907, and 0.868 for *Or22b*, *tjIIB*, and *dmrt93B*, respectively (all $P < 0.005$).

provide an attractive opportunity to study recent adaptation in a network context.

Copy number variation: The three large differentiated regions not located near centromeres (the *ace* region and the two chorion clusters; see above) have an interesting feature in common: most probes in each region have higher mean intensities in one population. For example, the chorion cluster on 3L contains 1106 probes in the 58-kb significant region: 91% of these probes have higher mean intensity in northern Australia than in southern Australia. The simplest explanation for this result is copy number variation—either a duplication with respect to the reference sequence is present at higher frequency in northern Australia or a deletion with respect to the reference sequence is at higher frequency in southern Australia. We scanned the genome for CNVs by recording the sign of the intensity difference for each probe and comparing this distribution to a null expectation from permuted data (see MATERIALS AND METHODS). We found many differentiated regions in which most probes (45 probes in a 50-probe window or 80 probes in a 100-probe window; see MATERIALS AND METHODS) had a higher mean intensity in one population, consistent with copy number variation. In the United States, 26 (4.7%) regions significant at the FDR < 0.05 threshold met this stringent criterion. In Australia, 22.3% of our significantly differentiated regions were also significant in this copy number variant analysis. However, other biological or technical factors could generate autocorrelation in mean probe intensity along chromosomes. Consider, for example, a region with many differentiated SNPs and strong linkage disequilibrium. If the reference strain matches a haplotype occurring at higher frequency in the south, then all SNP-containing probes would be more intense in the south. Extrachromosomal DNA or differential degrees of chromosome polytenization in two populations could also generate autocorrelated patterns of probe intensity.

For probes not containing a SNP, technical factors might create slight differences in mean intensity between populations that are correlated along chromosomes. For example, the nonrandom spatial distribution of GC content across Affymetrix arrays, combined with the existence of spatial artifacts not entirely corrected in our normalization, could result in nondifferentiated probes with spatially correlated intensities (for nondifferentiated probes, all variance is technical variance, so even if this variance is very small it will affect this test). Despite these caveats, it seems very likely that some fraction of our candidate CNVs is real and that the number of geographically differentiated CNVs in the genome is considerable.

To begin investigating these candidate CNVs, we used qPCR (see MATERIALS AND METHODS) to confirm that the largest differentiated candidate CNV, which spans nine genes and >40 kb, is a polymorphic duplication at higher frequency in northern Australia than in southern Australia ($P = 0.01$). This includes *ace*, which is also contained within a candidate CNV in the United States. However, the candidate U. S. CNV does not include the other genes, suggesting the possibility that Australian and U. S. CNVs originated independently (LABBE *et al.* 2007). We also validated, using PCR, that the single most significant euchromatic region in Australia is a 2098-bp deletion (Figure 6). This deletion, which includes portions of two olfactory receptors, *Or22a* and *Or22b*, is present in 9 of 12 northern chromosomes sequenced but in 0 of 11 southern chromosomes sequenced. It may have resulted from unequal crossing over between the 5' regions of these two genes, which are 95% identical. In deleted chromosomes, the first exon and intron of *Or22a* remain, along with exons 2–4 of *Or22b*. The deleted chromosomes therefore contain a potentially functional chimeric receptor. In addition to the deletion, there are five nonsynonymous SNPs, all of which are polymorphic in the north and monomorphic in the

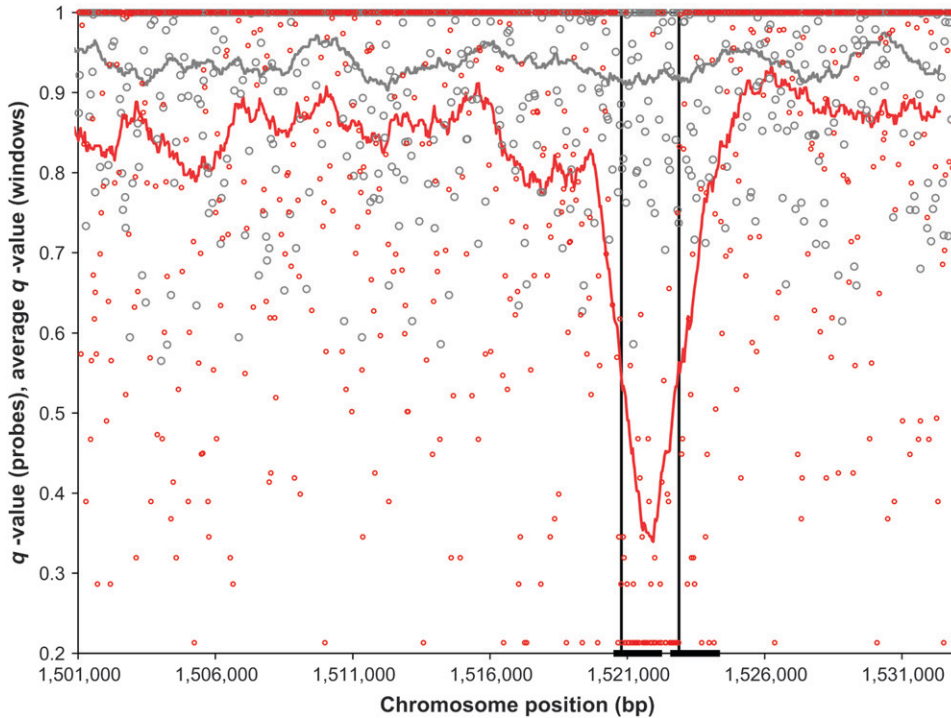


FIGURE 6.—Differentiation at *Or22a* and *Or22b*. Australian data are in red, the U.S. data are in gray, and the genes are boxes. Circles and dots denote individual probes, and the running averages of 50-probe windows are also shown. We verified the presence of a deletion that includes all but the first exon of *Or22a* and only the first exon of *Or22b*; the breakpoints of this deletion are indicated by vertical black lines. Most probes have q -values of 1, except in the deletion in Australia, where most probes have the lowest q -value, 0.213.

south; several are nearly fixed differences, with frequencies of 0% in the south and 95% in the north. Finally, two SNPs and a 16-bp insertion in the 3'-UTR are present in 85–90% of northern chromosomes but absent from southern chromosomes. *Or22a* and *Or22b* are coexpressed in the ab3A neuron, which is sexually dimorphic and responds to a broad range of odors (DOBRIŤSA *et al.* 2003). Functional analyses indicate that *Or22a* accounts for the full odor response of this neuron, with no obvious role of *Or22b* (DOBRIŤSA *et al.* 2003). Functional investigation of the differentiated alleles detected in this experiment may shed light on the potential selective forces involved in creating such strong differentiation at this locus in Australia.

Several candidate CNVs overlap previously identified CNVs: *meth8* overlaps a polymorphic deletion (A. D. KERN and D. J. BEGUN, unpublished results) and the differentiated CNV shared between continents at *cyp6a17* is coincident with a known polymorphic duplication (C. ROBIN, personal communication). The breakpoints of *In(3R)P* are candidate CNVs in our data, and these regions are known to be associated with duplications and deletions (MATZKIN *et al.* 2005). These data support the idea that many CNVs are influenced by spatially varying selection in *D. melanogaster*. A complete list of candidate CNVs is available as supplemental data.

Heterochromatin: We also analyzed heterochromatic regions of the version 5.0 genome assembly. These include the assembled portions of the chromosome 2, 3, and X centromeres, which are divided into right and left arms for chromosomes 2 and 3 (*e.g.*, *2Rhet*, *2Lhet*), and the entire fourth “dot” chromosome. These regions are large, gene poor, located near centromeres,

experience little crossing over, and harbor low levels of sequence variation within populations. However, despite the low levels of variation, our data suggest that these regions contain functionally important variants at different frequency in northern and southern populations. *2Rhet*, *3Lhet*, and *3Rhet* are very differentiated on both continents (Table 1). The fourth chromosome shows extreme differentiation in Australia, with 62% of the chromosome significant at 0.001 FDR, while the United States shows no evidence of differentiation on the chromosome. Overall, these data indicate that parts of the centromeres of chromosomes 2 and 3 are very differentiated between temperate and subtropical populations, while conclusions regarding the centromere of the X are less clear, with only one small significant region in Australia (Table 1). Although we used only probes with one perfect match in the reference genome, these probes may still be repetitive in nature, either because they are misassembled in the current reference or because the reference sequence happens to possess only one copy. Indeed, all of the significant regions in the centromeric heterochromatin are also significant as candidate copy number variants (although it is even more likely that we could mistake differentiated SNP haplotypes for CNVs in regions with high linkage disequilibrium). Functionally important heterochromatic variants could be associated with heterochromatic genes. Alternatively, the variation documented here could reflect functional differences among natural *D. melanogaster* centromeres, perhaps associated with variation in meiotic segregation (NOVITSKI 1955).

Conclusion: The list of differentiated probes and genomic regions presented here adds considerably to

the list of candidate genes harboring adaptive polymorphisms in *D. melanogaster*. Most regions that are differentiated in parallel on both continents are likely influenced by spatially varying selection. The dramatic overrepresentation of some functionally related groups of genes in differentiated regions implicates selection acting on these loci and functions as well. These data provide an excellent opportunity to study how sets of related genes adapt to the environment (including their interactions with each other). They also provide information that may be useful in refining the annotations of known elements and dissecting the functions of unknown elements in the genome. In addition to considerable population genetic and functional analysis motivated by the array data, much descriptive work remains to be done. For example, a substantial portion of the differentiation between populations may be due to copy number variants, but we did not fully characterize this variation. We are also uncertain about whether differences between continents reflect biological differences (different selection pressures or different adaptive responses to the same selection pressures) or differences in population genetics parameters (*e.g.*, the physical scale of linkage disequilibrium) that make regions detectable on only one continent. This is an especially interesting question, as some of our most dramatic differentiation was found only in Australia (*e.g.*, *Or22*, chorion genes, cytoskeleton genes, and the entire fourth chromosome).

We are especially grateful to Ary Hoffman and Paul Schmidt for sharing fly stocks. We also thank Sergey Nuzhdin, Alisha Holloway, Chuck Langley, and many others for advice and support. Finally, we thank Umbreen Arshad, Liz Milano, Matt Rolston, and Heather Lindfors, and others in the Begun lab for technical and laboratory assistance. This work was supported by seed money from the Center for Population Biology at the University of California, Davis, by National Science Foundation predoctoral fellowships to T.L.T. and M.T.L., and by National Institutes of Health grant GM071926 to D.J.B.

LITERATURE CITED

- ABDU, U., A. GONZALEZ-REYES, A. GHABRIAL and T. SCHUPBACH, 2003 The *Drosophila spn-D* gene encodes a RAD51C-like protein that is required exclusively during meiosis. *Genetics* **165**: 197–204.
- AGIS, M., and C. SCHLOTTERER, 2001 Microsatellite variation in natural *Drosophila melanogaster* populations from New South Wales (Australia) and Tasmania. *Mol. Ecol.* **10**: 1197–1205.
- ANDERSON, A. R., J. E. COLLINGE, A. A. HOFFMANN, M. KELLETT and S. W. McKECHNIE, 2003 Thermal tolerance trade-offs associated with the right arm of chromosome 3 and marked by the *hsp-omega* gene in *Drosophila melanogaster*. *Heredity* **90**: 195–202.
- ANDERSON, A. R., A. A. HOFFMANN, S. W. McKECHNIE, P. A. UMINA and A. R. WEEKS, 2005 The latitudinal cline in the *In(3R)Payne* inversion polymorphism has shifted in the last 20 years in Australian *Drosophila melanogaster* populations. *Mol. Ecol.* **14**: 851–858.
- ANDOLFATTO, P., and M. KREITMAN, 2000 Molecular variation at the *In(2L)t* proximal breakpoint site in natural populations of *Drosophila melanogaster* and *D. simulans*. *Genetics* **154**: 1681–1691.
- ANDOLFATTO, P., and J. D. WALL, 2003 Linkage disequilibrium patterns across a recombination gradient in African *Drosophila melanogaster*. *Genetics* **165**: 1289–1305.
- BARTON, N. H., 1983 Multilocus clines. *Evolution* **37**: 454–471.
- BERRY, A., and M. KREITMAN, 1993 Molecular analysis of an allozyme cline: alcohol-dehydrogenase in *Drosophila melanogaster* on the east coast of North America. *Genetics* **134**: 869–893.
- BETTENCOURT, B. R., I. KIM, A. A. HOFFMANN and M. E. FEDER, 2002 Response to natural and laboratory selection at the *Drosophila hsp70* genes. *Evolution* **56**: 1796–1801.
- BOREVITZ, J., D. LIANG, D. PLOUFFE, H.-S. CHANG, T. ZHU *et al.*, 2003 Large scale identification of single feature polymorphisms in complex genomes. *Genome Res.* **13**: 513–523.
- CALBOLI, F. C. F., W. J. KENNINGTON and L. PARTRIDGE, 2003 QTL mapping reveals a striking coincidence in the positions of genomic regions associated with adaptive variation in body size in parallel clines of *Drosophila melanogaster* on different continents. *Evolution* **57**: 2653–2658.
- CARPENTER, A. T. C., 1994 Egalitarian and the choice of cell fates in *Drosophila melanogaster* oogenesis, pp. 223–246 in *Germline Development*, edited by J. MARSH and J. GOODE. Wiley, New York.
- CHUNG, H., M. R. BOGWITZ, C. McCART, A. ANDRIANOPOULOS, R. H. FRENCH-CONSTANT *et al.*, 2007 *Cis*-regulatory elements in the Accord retrotransposon result in tissue-specific expression of the *Drosophila melanogaster* insecticide resistance gene *Cyp6g1*. *Genetics* **175**: 1071–1107.
- COPPIN, C. W., W. A. ODGERS and J. G. OAKESHOTT, 2007 Latitudinal clines for nucleotide polymorphisms in the *Esterase-6* gene of *Drosophila melanogaster*. *Genetica* **129**: 259–271.
- DABORN, P. J., C. LUMB, A. BOEY, W. WONG, R. H. FRENCH-CONSTANT *et al.*, 2007 Evaluating the insecticide resistance potential of eight *Drosophila melanogaster* cytochrome P450 genes by transgenic over expression. *Insect Biochem. Mol. Biol.* **37**: 512–519.
- DE JONG, G., and Z. BOCHDANOVITS, 2003 Latitudinal clines in *Drosophila melanogaster*: body size, allozyme frequencies, inversion frequencies, and the insulin-signalling pathway. *J. Genet.* **82**: 207–223.
- DOBRTSA, A. A., W. VAN DER GOES VAN NATERS, C. G. WARR, R. A. STEINBRECHT and J. R. CARLSON, 2003 Integrating the molecular and cellular basis of odor coding in the *Drosophila* antenna. *Neuron* **37**: 827–841.
- EANES, W. F., 1999 Analysis of selection on enzyme polymorphisms. *Annu. Rev. Ecol. Syst.* **30**: 301–326.
- ENDLER, J. A., 1977 *Geographic Variation, Speciation and Clines*. Princeton University Press, Princeton, NJ.
- FELSENSTEIN, J., 1976 The theoretical population genetics of variable selection and migration. *Annu. Rev. Genet.* **10**: 253–280.
- FIUMERA, A. C., B. L. DUMONT and A. G. CLARK, 2005 Sperm competitive ability in *Drosophila melanogaster* associated with variation in male reproductive proteins. *Genetics* **169**: 243–257.
- FOURNIER, D., F. KARCH, J. M. BRIDE, L. M. C. HALL, J. B. BERGE *et al.*, 1989 *Drosophila melanogaster acetylcholinesterase* gene: structure, evolution and mutations. *J. Mol. Biol.* **210**: 15–22.
- GAUTIER, L., L. COPE, B. M. BOLSTAD and R. A. IRIZARRY, 2004 affy: analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics* **20**: 307–315.
- 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.
- GOCKEL, J., S. J. W. ROBINSON, W. J. KENNINGTON, D. B. GOLDSTEIN and L. PARTRIDGE, 2002 Quantitative genetic analysis of natural variation in body size in *Drosophila melanogaster*. *Heredity* **89**: 145–153.
- GRESHAM, D., D. M. RUDERFER, S. C. PRATT, J. SCHACHERER, M. DUNHAM *et al.*, 2006 Genome-wide detection of polymorphism at nucleotide resolution with a single DNA microarray. *Science* **311**: 1932–1936.
- HADRILL, P. R., B. CHARLESWORTH, D. L. HALLIGAN and P. ANDOLFATTO, 2005 Patterns of intron sequence evolution in *Drosophila* are dependent upon length and GC content. *Genome Biol.* **6**: r67.
- HALE, L. R., and R. S. SINGH, 1991 A comprehensive study of genic variation in natural populations of *Drosophila melanogaster*. IV. Mitochondrial DNA variation and the role of history *vs.* selection in the genetic structure of geographic populations. *Genetics* **129**: 103–117.
- HOFFMANN, A. A., A. ANDERSON and R. HALLAS, 2002 Opposing clines for high and low temperature resistance in *Drosophila melanogaster*. *Ecol. Lett.* **5**: 614–618.

- HUDSON, R. R., M. SLATKIN and W. P. MADDISON, 1992 Estimation of levels of gene flow from DNA sequence data. *Genetics* **132**: 583–589.
- JAMES, A. C., and L. PARTRIDGE, 1995 Thermal evolution of rate of larval development in *Drosophila melanogaster* in laboratory and field populations. *J. Evol. Biol.* **8**: 315–330.
- KAWAGUCHI, S., and Y. ZHENG, 2004 Characterization of a *Drosophila* centromere protein CP309 that shares homology with kendrin and CG-NAP. *Mol. Biol. Cell* **15**: 37–45.
- 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.
- KENNINGTON, W. J., L. PARTRIDGE and A. A. HOFFMANN, 2006 Patterns of diversity and linkage disequilibrium within the cosmopolitan inversion *In(3R)Payne* in *Drosophila melanogaster* are indicative of coadaptation. *Genetics* **172**: 1655–1663.
- KENNINGTON, W. J., A. A. HOFFMANN and L. PARTRIDGE, 2007 Mapping regions within cosmopolitan inversion *In(3R)Payne* associated with natural variation in body size in *Drosophila melanogaster*. *Genetics* **177**: 549–556.
- KNIBB, W. R., 1982 Chromosome inversion polymorphisms in *Drosophila melanogaster*. 2. Geographic clines and climatic associations in Australasia, North America and Asia. *Genetica* **58**: 213–221.
- LABBE, P., A. BERTHOMIEU, C. BERTICAT, H. ALOUT, M. RAYMOND *et al.*, 2007 Independent duplications of the *acetylcholinesterase* gene conferring insecticide resistance in the mosquito *Culex pipiens*. *Mol. Biol. Evol.* **24**: 1056–1067.
- LACHAISE, D., and J. F. SILVAIN, 2004 How two Afrotropical endemics made two cosmopolitan human commensals: the *Drosophila melanogaster*-*D. simulans* palaeogeographic riddle. *Genetica* **120**: 17–39.
- LAROCQUE, J. R., B. JAKLEVIC, T. T. SU and J. SEKELSKY, 2007 *Drosophila* ATR in double-strand break repair. *Genetics* **175**: 1023–1033.
- 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.
- LIVAK, K. J., and T. D. SCHMITTGEN, 2001 Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)). *Methods* **25**: 402–408.
- LONG, A. D., and R. S. SINGH, 1995 Molecules versus morphology: the detection of selection acting on morphological characters along a cline in *Drosophila melanogaster*. *Heredity* **74**: 569–581.
- MARTINEZ-CAMPOS, M., R. BASTO, J. BAKER, M. KERNAN and J. W. RAFF, 2004 The *Drosophila* pericentrin-like protein is essential for cilia/flagella function, but appears to be dispensable for mitosis. *J. Cell Biol.* **165**: 673–683.
- MATTHIES, H. J. G., L. G. MESSINA, R. NAMBA, K. J. GREER, M. Y. WALKER *et al.*, 1999 Mutations in the *alpha-tubulin 67C* gene specifically impair achiasmate segregation in *Drosophila melanogaster*. *J. Cell Biol.* **147**: 1137–1144.
- MATZKIN, L. M., T. J. S. MERRITT, C. T. ZHU and W. F. EANES, 2005 The structure and population genetics of the breakpoints associated with the cosmopolitan chromosomal inversion *In(3R)Payne* in *Drosophila melanogaster*. *Genetics* **170**: 1143–1152.
- NORRY, F. M., J. DAHLGAARD and V. LOESCHCKE, 2004 Quantitative trait loci affecting knockdown resistance to high temperature in *Drosophila melanogaster*. *Mol. Ecol.* **13**: 3585–3594.
- NOVITSKI, E., 1955 Genetic measures of centromere activity in *Drosophila melanogaster*. *J. Cell. Comp. Physiol.* **45**: 151–169.
- OAKESHOTT, J. G., J. B. GIBSON, P. R. ANDERSON, W. R. KNIBB, D. G. ANDERSON *et al.*, 1982 Alcohol dehydrogenase and glycerol-3-phosphate dehydrogenase clines in *Drosophila melanogaster* on different continents. *Evolution* **36**: 86–96.
- OIKEMUS, S. R., J. QUEIROZ-MACHADO, K. LAI, N. MCGINNIS, C. SUNKEL *et al.*, 2006 Epigenetic telomere protection by *Drosophila* DNA damage response pathways. *PLoS Genet.* **2**: 693–706.
- OMETTO, L., S. GLINKA, D. DE LORENZO and W. STEPHAN, 2005 Inferring the effects of demography and selection on *Drosophila melanogaster* populations from a chromosome-wide scan of DNA variation. *Mol. Biol. Evol.* **22**: 2119–2130.
- ROZAS, J., J. C. SANCHEZ-DELBARRIO, X. MESSEGUER and R. ROZAS, 2003 DnaSP, DNA polymorphism analyses by the coalescent and other methods. *Bioinformatics* **19**: 2496–2497.
- SAWYER, L. A., J. M. HENNESSY, A. A. PEIXOTO, E. ROSATO, H. PARKINSON *et al.*, 1997 Natural variation in a *Drosophila* clock gene and temperature compensation. *Science* **278**: 2117–2120.
- SCHMIDT, P. S., A. B. PAABY and A. S. HESCHEL, 2005 Genetic variance for diapause expression and associated life histories in *Drosophila melanogaster*. *Evolution* **59**: 2616–2625.
- SEZGIN, E., D. D. DUVERNELL, L. M. MATZKIN, Y. H. 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.
- 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.
- STALKER, H. D., 1976 Chromosome studies in wild populations of *D. melanogaster*. *Genetics* **82**: 323–347.
- STEFFENSEN, S., P. A. COELHO, N. COBBE, S. VASS, M. COSTA *et al.*, 2001 A role for *Drosophila* SMC4 in the resolution of sister chromatids in mitosis. *Curr. Biol.* **11**: 295–307.
- STOREY, J. D., and R. TIBSHIRANI, 2003 Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci. USA* **100**: 9440–9445.
- TURNER, T. L., M. W. HAHN and S. V. NUZHIDIN, 2005 Genomic islands of speciation in *Anopheles gambiae*. *PLoS Biol.* **3**: e285.
- WEILL, M., P. FORT, A. BERTHOMIEU, M. P. DUBOIS, N. PASTEUR *et al.*, 2002 A novel *acetylcholinesterase* gene in mosquitoes codes for the insecticide target and is non-homologous to the *ace* gene in *Drosophila*. *Proc. R. Soc. Lond. Ser. B Biol. Sci.* **269**: 2007–2016.
- WEILL, M., G. LUTFALLA, K. MOGENSEN, F. CHANDRE, A. BERTHOMIEU *et al.*, 2003 Insecticide resistance in mosquito vectors. *Nature* **423**: 136–137.
- WERNER, J. D., J. O. BOREVITZ, N. WARTHMAN, G. T. TRAINER, J. R. ECKER *et al.*, 2005 Quantitative trait locus mapping and DNA array hybridization identify an FLM deletion as a cause for natural flowering-time variation. *Proc. Natl. Acad. Sci. USA* **102**: 2460–2465.
- WINZELER, E. A., D. R. RICHARDS, A. R. CONWAY, A. L. GOLDSTEIN, S. KALMAN *et al.*, 1998 Direct allelic variation scanning of the yeast genome. *Science* **281**: 1194–1197.
- ZHANG, L., M. F. MILES and K. D. ALDAPE, 2003 A model of molecular interactions on short oligonucleotide microarrays. *Nat. Biotechnol.* **21**: 818–821.
- ZHU, L., J. WILKEN, N. B. PHILLIPS, U. NARENDRA, G. CHAN *et al.*, 2000 Sexual dimorphism in diverse metazoans is regulated by a novel class of intertwined zinc fingers. *Genes Dev.* **14**: 1750–1764.