DOI: 10.1534/genetics.104.026252

Nucleotide Variation in the *Egfr* Locus of *Drosophila melanogaster*

Arnar Palsson,[1,2] Ann Rouse,[1,3] Rebecca Riley-Berger, Ian Dworkin and Greg Gibson[4]

Department of Genetics, North Carolina State University, Raleigh, North Carolina 27513-7614

Manuscript received January 6, 2004
Accepted for publication April 7, 2004

ABSTRACT

The *Epidermal growth factor receptor* is an essential gene with diverse pleiotropic roles in development throughout the animal kingdom. Analysis of sequence diversity in 10.9 kb covering the complete coding region and 6.4 kb of potential regulatory regions in a sample of 250 alleles from three populations of *Drosophila melanogaster* suggests that the intensity of different population genetic forces varies along the locus. A total of 238 independent common SNPs and 20 indel polymorphisms were detected, with just six common replacements affecting >1475 amino acids, four of which are in the short alternate first exon. Sequence diversity is lowest in a 2-kb portion of intron 2, which is also highly conserved in comparison with *D. simulans* and *D. pseudoobscura*. Linkage disequilibrium decays to background levels within 500 bp of most sites, so haplotypes are generally restricted to up to 5 polymorphisms. The two North American samples from North Carolina and California have diverged in allele frequency at a handful of individual SNPs, but a Kenyan sample is both more divergent and more polymorphic. The effect of sample size on inference of the roles of population structure, uneven recombination, and weak selection in patterning nucleotide variation in the locus is discussed.

THE *Epidermal growth factor receptor* (*Egfr*) in Drosophila is involved in and quite essential for a wide range of activities, but there is little indication that it has ever had a central role in promoting evolutionary change. Unlike the *Hox* family, for example, which is widely recognized as a key mediator of morphological diversification (Gellon and McGinnis 1998), or numerous enzymes in central metabolism that facilitate ecological adaptation (Eanes 1999), *Egfr* is more obviously regarded as a generic signaling component that likely helps to promote developmental stability. In the absence of compelling implication of involvement in some unusual phenomenon, it is the sort of gene that escapes the notice of evolutionary geneticists. Yet to the extent that quantitative variation is the product of subtle polymorphisms in hundreds of loci, it is also arguably the sort of gene that should receive attention.

Point estimation of the basic parameters of nucleotide variation is generally recognized to be unaffected by sample size, and since the tests of neutrality based on these estimates have low power to detect weak selection, most population genetic surveys deal with samples of no more than 30 alleles. However, our study was motivated primarily by quantitative genetic questions, requiring that we have confidence in estimates of the level and variance of linkage disequilibrium—and hence haplotype structure—throughout the locus, and assessment of the presence or absence of population structure. As shown in the accompanying article (Palsson and Gibson 2004, accompanying article in this issue), it turns out that complete sequence coverage rather than selective genotyping is critical to the detection of genotype-phenotype associations. We argue here that for some purposes sample sizes of several hundred alleles are likely to yield information of interest to molecular evolutionists that cannot be gleaned from smaller samples. In particular, significance testing of population structure, inference that recombination rates may vary along a locus, and monitoring of the frequency of rare and complex alleles all depend on larger sample sizes.

We chose *Egfr* over thousands of other typical loci for two main reasons. First, it has highly conserved roles in cell growth and differentiation throughout the animal kingdom (Freeman 1998; Stein and Staros 2000; Shilo 2003), so is intrinsically interesting in terms of understanding the genetics of developmental pathways. Second, it is an excellent candidate gene for quantitative association with a range of traits in Drosophila (Dworkin *et al.* 2003; Palsson and Gibson 2004, accompanying article in this issue) yet has so many pleiotropic functions (*e.g.*, Clifford and Schüpbach 1994; Wang *et al.* 2000; Duchek and Rorth 2001; Galindo *et al.* 2002; Johnson-Hamlet and Perkins 2002; Zecca and Struhl 2002; Yang and Baker 2003) that it is easy to imagine that very little variation would be tolerated in

[1] These authors contributed equally to this work.

[2] *Present address:* Department of Ecology and Evolution, University of Chicago, Chicago, IL 60637.

[3] *Present address:* Department of Biology, Duke University, Durham, NC 27708.

[4] *Corresponding author:* Department of Genetics, Gardner Hall, North Carolina State University, Raleigh, NC 27695-7614. E-mail: ggibson@unity.ncsu.edu

the protein structure, as seems to be the case. These different functions could also set up conflicting trade-offs at different stages of development, and linkage disequilibrium between regulatory sites affecting different domains of expression could impact polymorphism, perhaps resulting in strong balancing selection. Similar to downstream signaling components (Riley *et al.* 2003), we do not find evidence for the latter, but discuss some unexpected haplotype structure and population differentiation in the context of weak selection, uneven recombination, and a predominance of purifying selection.

MATERIALS AND METHODS

Drosophila stocks: Thirty-six Kenyan lines were established from stocks obtained from the Bowling Green Stock Center in 1997, based on flies collected by R. Woodruff in the 1970s. Second chromosomes of these lines were isogenized by passage over the CyO balancer in the Samarkand background. Several cases of probable gene conversion to the CyO allele were observed upon sequencing, but these were excluded from further analysis. Eighty-three Californian isofemale lines were collected by S. Nuzhdin from the Wolfskill orchard near Winters, California (CA), in 1998, and subjected to >40 generations of pairwise sib mating to generate the near-isogenic lines surveyed here (Yang and Nuzhdin 2003). Similarly, 150 North Carolinian near-isogenic lines were established from isofemales trapped by us from several bins in a peach orchard near West End, North Carolina (NC), in July 2000. After 10 generations of sib mating only 70% of the *Egfr* loci in these lines were homozygous, so sib mating was continued in multiple sublines for a further 5 generations, and homozygous lines were then chosen on the basis of sequencing. A total of 130 NC near-isogenic lines survived, 15 of which remained at least partially heterozygous (*cf.* 2 of the CA lines), in which case that portion of the locus was ignored. It is probable that the process of inbreeding and selection of homozygous chromosomes biases the sampling, but no indication of departure from Hardy-Weinberg proportions was observed in the data and if there is a bias in the data set, it is toward an excess of rare alleles. The *Drosophila simulans* allele was obtained from the same population of West End flies. Flies were maintained in vials on 10 ml standard cornmeal medium supplemented with yeast.

Sequencing: All single-nucleotide polymorphism (SNP) and insertion/deletion (indel) genotypes were obtained by direct sequencing of PCR products obtained from squish-preps of a single male fly per line. The 10.9 kb was surveyed with 6 amplification reactions, ranging from 1.2 to 2.1 kb in length, and 17 sequencing reactions generated from the external PCR primers and an internal primer (listed in supplementary Table 1 at http://statgen.ncsu.edu/ggibson/supplinfo/supplinfo6.htm) for most of the fragments, as diagrammed in Figure 1. Amplification reactions were performed with Promega (Madison, WI) *Taq* polymerase, products were purified from agarose gels with QIAquick columns (QIAGEN, Valencia, CA), and sequencing was performed with Big Dye II mix (Perkin-Elmer, Norwalk, CT) on ABI 3700 sequencers at the North Carolina State University Genome Research Laboratory. Most high-quality reads were in the range of 650–700 bases, but some regions proved particularly refractory to both amplification and sequencing, due mainly to high SNP and indel polymorphism and to homopolymeric runs. New primer combinations were tested, particularly in the vicinity of exons 1 and 2, with limited success, and consequently depth of coverage varies along the locus. No attempt was made to sequence all fragments on both strands, due to a trade-off in time and expense against gain in accuracy. However, we estimate the base-calling error rate at <0.1% per polymorphic site on the basis of repeat sequencing of some alleles and up to 100-bp overlaps between ends of sequence fragments that resulted in ~1.5× coverage with 18% of all nucleotides represented in at least two traces. There is slightly >2 Mb of completed sequence, and just four unresolved polymorphisms remain after manual inspection and comparison of pairs of traces.

Nevertheless, the genotype matrix for the entire collection of 257 partial *Egfr* alleles is only 74% complete, with an average read length of 8067 (±131) bp. The three main reasons for incomplete coverage are (i) reduced sampling of problematic regions, most notably around exon 2 in the CA lines; (ii) lingering heterozygosity in portions of some alleles, primarily in the NC lines; and (iii) reduced finishing effort for the Kenyan lines, as these were of less interest to us in relation to our genotype-phenotype mapping studies. Consequently all molecular evolutionary conclusions should be interpreted with these caveats of possible sequence error and sampling bias in mind. Locations of variants are specified relative to GenBank entry 17571116 (also available as supplementary information at http://statgen.ncsu.edu/ggibson/SupplInfo/SupplInfo6.htm), which spans 48 kb encompassing the complete *Egfr* locus and up to 5 kb on either end. Coverage of the predicted coding region in our sequence is 100%; the extent of *Egfr* regulatory sequences is unknown, but at least one-third of the 24-kb intron 1 is occupied by three other open reading frames (ORFs), and the flanking genes are closely adjacent, so coverage of the entire locus lies between 25 and 75%.

All polymorphisms were aligned manually and inspected by two observers, and only unambiguous SNPs are reported. Trace files were imported into the ContigExpress module of VectorNTI Version 5 (Informax) for primary editing and assembly of contiguous alleles for each region from overlapping sequence reads. Sequence alignment was completed using ClustalW (Thompson *et al.* 1994), and the matrix of alleles was transferred to Genedoc (Nicholas *et al.* 1997), which facilitates manual adjustment of indel polymorphism. Final sequences are available from GenBank: *D. melanogaster* alleles are divided by exons and populations as follows: exon 1 [CA, AY460697–AY460774; Kenya (K), AY460775–AY460801; NC, AY460802–AY460927], exon 2 (CA, AY460928–AY460979; K, AY460980–AY461000; NC, AY461001–AY461101), and exons 3–6 (CA, AY461102–AY461184; K, AY461185–AY461220; NC, AY461221–AY461357). Outgroup sequences are from *D. simulans*, for exons 1 (AY460690–AY460692), 2 (AY460693), and 3–6 (AY460694), in addition to one *D. sechellia* allele for exon 1 (AY460689).

Analysis of nucleotide diversity: Since the Kenyan sample had a large number of unconfirmed singletons and is less than one-fifth the size of the North American sample, all analyses aside from the F_{ST} comparison are based on the combined NC and CA sample. Nucleotide diversity was initially analyzed as the pairwise distance between alleles (π) and on the basis of the number of segregating sites (θ) in both DnaSP Version 3.53 (Rozas and Rozas 1999) and Tassel Version 2.0 (E. Buckler, http://www.maizegenetics.org). Inconsistencies between these estimates around exons 1 and 2 arose as a result of the way the two programs handle missing data, so we report more direct manual estimates. Nucleotide diversity, π, was estimated as the per-nucleotide expected heterozygosity in a sliding window of 100 bp in Figure 2A and over each functional region of the gene in Table 1. The estimates of θ per nucleotide in Table 1 were obtained as S/na_i, where S is the observed

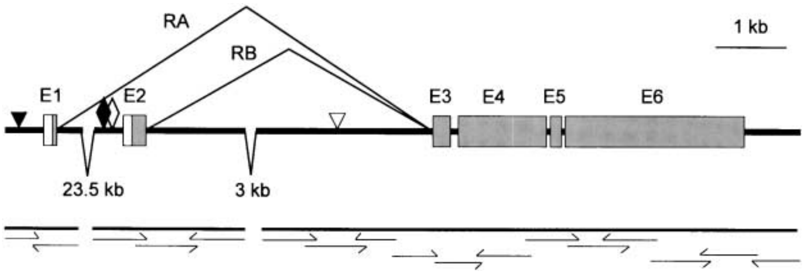

FIGURE 1.—Structure of the *Egfr* locus and sequencing strategy. The six exons are indicated as shaded boxes, with open boxes preceding exons 1 and 2 representing putative signal peptides. Alternative splicing results in two transcripts, RA and RB, in which alternate first exons are joined to the four common 3′ exons. The first two introns are 24.4 and 6.9 kb in length, respectively, and the extent of unsequenced DNA (including three ORFs on the opposite strand, not shown) is indicated in kilobases. Three noteworthy sequence features mentioned in the text are a CAA microsatellite (open diamond) and an unusual CN repeat (solid diamond) upstream of exon 2, and a conserved 30-bp motif in intron 2 (open triangle). A short *pogo* element insertion was found in two of the NC alleles, 253 bp upstream of the start codon (solid triangle). The sequencing strategy shown below the gene resulted in three contigs assembled from six PCR fragments and 17 sequencing reactions, with half-arrows indicating direction of sequence extension.

number of segregating sites, and a_i is the sum over $i = 1$ to $N - 1$ of $1/i$ for N, the mean number of alleles in the segment of length n nucleotides (WATTERSON 1975). These two estimates track closely and while there is a tendency for $\theta > \pi$, this is not significant in any region or across the entire locus. Tests of neutrality (HUDSON *et al.* 1987; TAJIMA 1989; FU and LI 1993; FAY and WU 2000) were implemented in DnaSP Version 3.53 on reduced data sets containing only complete sequences for sliding windows and/or specific segments of the locus, but did not provide any strong evidence for either directional or balancing selection: PALSSON (2003) provides a detailed description. The effects of sample size on tests of neutrality were determined by 100 jackknife iterations with $N = 12$, 48, or 96, for three regions of ~1000 bp each that had no missing data. Protein divergence (MCDONALD and KREITMAN 1991) was assessed with Fisher's exact test. Alignment with the *D. pseudoobscura* homolog (raw contig 1071 downloaded from http://www.hgsc.bcm.tmc.edu) was analyzed with AVID and visualized with VISTA (MAYOR *et al.* 2000).

Linkage disequilibrium (LD) was assessed in Tassel, using Fisher's exact test (LEWONTIN 1988; WEIR 1996) to assess the significance of the squared correlation in allele frequencies, r^2. All analyses were performed for NC and CA separately and as a combined data set, and results are reported for the combined North American data set since no qualitative differences in patterns of LD among populations were detected. Only sites that were present in at least 50 alleles were included in the analyses, and only if both variants were observed in 5 or more alleles. The second criterion is critical due to limitations of contingency tests with small marginal counts (UPTON 1982), and this has demonstrated effects for estimates of linkage disequilibrium (LEWONTIN 1995). The effects of sample size, minimum number of alleles, and rare variants were explored by jackknife procedures in Tassel (with missing data) and those of sample size only in DnaSP 3.53 (without missing data).

Population differentiation among all three populations was estimated using the AMOVA procedure in Arlequin Version 2.0 (SCHNEIDER *et al.* 2000) for pairs of populations. The significance of F_{ST} parameters was assessed by 10,000 permutations with Bonferroni adjustment for the 258 comparisons involving common SNPs and indels in all three populations. Initial analyses were performed with sliding windows of haplotypes of 10 or 5 sites, and subsequently for each site separately. Demonstration of the power gained with larger samples was demonstrated by 1000 jackknife samples of the top four most population-stratified sites.

Homogeneity of the distribution of indel events along the locus, by regions defined in Table 1, was analyzed with a goodness-of-fit test (SCHAEFFER 2002). The effect of sample size on this test and on the variance in indel size was also estimated by 1000 jackknife iterations, with a sample size of 122. The shape of the microsatellite length distribution was studied by fitting a normal distribution and assessing deviations from uniformity by a Kolmogorov-Smirnov test. Skewness and kurtosis of the observed distributions were also compared to estimates derived from permutations programmed in R, with individual repeats scrambled with respect to one another.

RESULTS

Parameters of molecular variation in *Egfr*: Almost 10.9 kb of *Egfr* was sequenced in three segments of 1250, 2090, and 7523 nucleotides, corresponding, respectively, to DNA including the first and second alternate 5′ exons and the cluster of four 3′ exons that encode the bulk of the protein. A schematic of the gene structure is shown in Figure 1, and parameters summarizing the polymorphism are provided in Table 1. A total of 523 biallelic SNPs, 24 triallelic SNPs, and 70 indels were detected, including 22 amino acid replacement polymorphisms. A Hudson-Kreitman-Aguadé test on intronic *vs.* exonic sequences was not significant ($P = 0.89$). Sequence diversity estimated on the basis of the number of segregating sites (θ) or as the average nucleotide heterozygosity (π) hovers around 0.01 substitutions per nucleotide, and there is an average of 1 SNP/20 bp. This is well within the normal range for *D. melanogaster* genes (MORIYAMA and POWELL 1996). The estimates of θ and π vary along the locus, but do not differ significantly from one another in any region, suggesting that to a broad approximation the *Egfr* is evolving according to the expectations of neutral theory subject to varying levels of constraint in regulatory, exonic, and intronic sequences. Sample size >20 sequences has little effect on point estimation of π, θ, or haplotype diversity, but various tests of neutrality give more heterogeneous results with smaller samples, as indicated in supplementary Table 3 at http://statgen.ncsu.edu/ggibson/SupplInfo/SupplInfo6.htm.

There is a low level of amino acid replacement polymorphism in the Drosophila EGF receptor (DER) pro-

TABLE 1

Parameters of nucleotide diversity in the Drosophila *EGFR*

		Segregating polymorphisms							*cf. D. simulans*		
Region	Length (bp)	Start[a]	Het[b]	θ[c]	S_{tot}[d]	S_{comm}	Repl (rare)[e]	Indel[f]	Silent	Repl	Indel
5′ exon 1	605	5402	0.015	0.013	46	23	—	7	13	—	7
Exon 1	154	6016	0.009	0.008	7	5	6 (2)	1	0	6	1
Intron 1	492	6170	0.007	0.008	21	9	—	7	ND	—	ND
5′ exon 2	416	30120	0.006	0.010	23	6	—	9	13	—	1
Exon 2	295	30518	0.008	0.013	20	6	3 (3)	0	5	0	0
3′exon 2	1,384	30819	0.012	0.011	82	42	—	16	27	—	8
Intron 2	2,425	35340	0.004	0.008	112	27	—	13	25	—	2
Exon 3	223	37757	0.015	0.010	13	10	0 (0)	0	6	0	0
Intron 3	170	37980	0.052	0.037	30	20	—	5	10	—	1
Exon 4	1,175	38116	0.010	0.008	54	32	1(1)	0	22	0	0
Intron 4	66	39291	0.012	0.015	6	2	—	3	6	—	0
Exon 5	133	39358	0.017	0.016	12	5	2 (1)	0	0	2	0
Intron 5	74	39491	0.009	0.012	5	1	—	1	7	—	1
Exon 6	2,446	39561	0.008	0.007	95	56	8 (8)	1	44	2	0
3′ UTR	347	42010	0.010	0.010	20	11	—	2	6	—	1
Intergenic	455	42355	0.003	0.005	13	4	—	0	9	—	0
Total	10,863		0.009	0.009	546	259	20 (15)	65	193	10	22

[a] The number of the first base of the region is GenBank Drosophila accession NG_000184 (17571116). Segment lengths do not necessarily match perfectly with our sequence lengths because of indel variation among sequences.

[b] Het (heterozygosity) estimated as the average expected nucleotide heterozygosity at Hardy-Weinberg proportions.

[c] θ per nucleotide, estimated as S_{tot} divided by the sum from $i = 1$ to n of $1/n$, for n, the average number of alleles in the region, as well as by the number of nucleotides in the sequence segment.

[d] S_{tot} is the total number of segregating SNPs, and S_{comm} is the number of common SNPs (less common allele >5%).

[e] Repl (rare) is the number of replacement (or rare replacement) polymorphisms.

[f] Indel information is for the North American sample only.

tein sequence. Only six "common" replacement polymorphisms (where both alleles have frequencies >0.05) are observed, and four of these are located within the putative signal peptide region of alternate exon 1. All of the remaining 15 nonsynonymous SNPs are present in just 1, 2, or 4 of the 210 sampled North American sequences as listed in supplementary Table 2 (http://statgen.ncsu.edu/ggibson/SupplInfo/SupplInfo6.htm), and half of these affect amino acids that are conserved in the *D. pseudoobscura* DER sequence. An exonic 9-bp deletion polymorphism that replaces a conserved Pro-Asn$_3$ with His in exon 6 of a Californian allele is likely to be deleterious. Two of the North Carolinian alleles appear to have one instead of two initiating methionine codons in alternate exon 1, but polarization with the sibling species indicates that the derived allele is the insertion of 3 bp that is approaching fixation in *D. melanogaster*. The ratio of replacement to synonymous substitutions in the entire coding region for the comparison with a *D. simulans* allele is 0.05, which is normal for Drosophila, but is generally taken as a sign of strong functional constraint on the protein (Aquadro *et al.* 2001).

Each region of *Egfr* tells a slightly different story concerning the possible evolutionary forces acting on the gene, when intraspecific variation (Figure 2A) is contrasted with divergence over short (Figure 2B, *D. simulans*, ∼2.5 million years) or long (Figure 2C, *D. pseudoobscura*, 45 million years) timescales (Powell 1996). In the next four paragraphs, we discuss the parameters of variation in the two alternate 5′ exons, in intron 2, and in the main coding region.

Exon 1 is remarkable for being considerably more polymorphic than alternate first exon 2 at the protein level. It encodes by far the most polymorphic 51-amino-acid segment of the DER protein, and there are also six replacements relative to the *D. simulans* allele compared with just four in the remaining 1424 residues. Noncoding variation is at a slightly higher level than that observed elsewhere in the locus, and there are 15 indel variants in just 1.1 kb flanking the exon. Oddly, there are no synonymous substitutions relative to *D. simulans*, but this is not sufficient to provide evidence for positive selection by the McDonald-Kreitman test (McDonald and Kreitman 1991; two-tailed $P = 0.068$, Fisher's exact test). Comparison with *D. pseudoobscura* also indicates relaxed constraint: unlike the remainder of the locus, only a few short stretches of >50% nucleotide identity are seen, and apart from a putative start site and splice donor sites, alignment was impossible.

Exon 2 is characterized by an unusually low level of haplotype structure and by a slight but nonsignificant

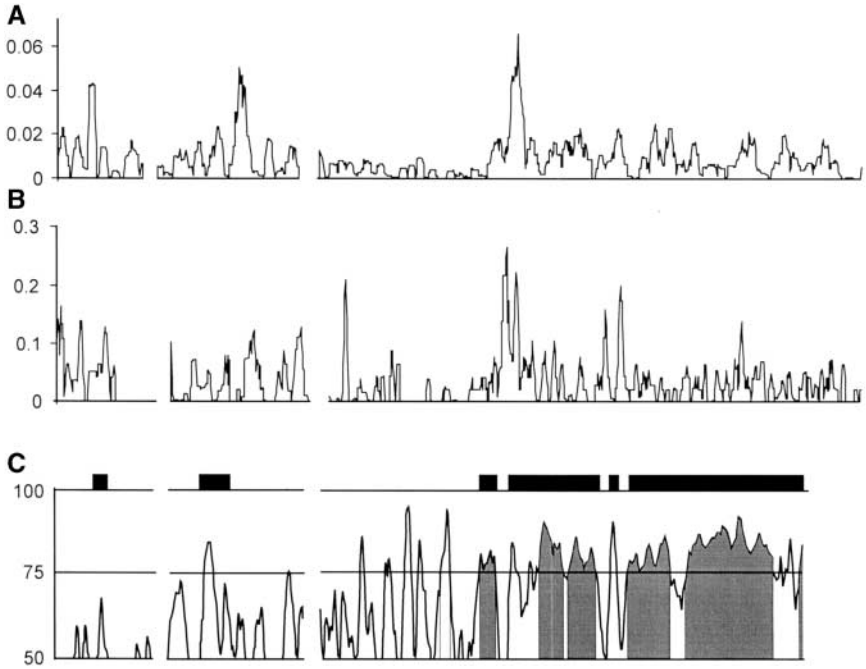

Figure 2.—Polymorphism and divergence. (A) Sequence diversity within *D. melanogaster* in the North American populations is plotted as heterozygosity per nucleotide in a 100-bp sliding window from 5′ to 3′ with two breaks between contigs. (B) Sequence divergence relative to a *D. simulans* allele sampled from the NC collection site: note the peaks of divergence in introns 3, 4, and 5, whereas only intron 3 is highly polymorphic within *D. melanogaster*. (C) VISTA plot of sequence conservation (percentage of nucleotide identity) relative to *D. pseudoobscura*, with 100-bp stretches >75% identical shaded. Solid bars at the top show location of exons. Note that peaks of conservation correspond broadly with conserved regions in the *D. melanogaster* group, but that only exon 1 is very poorly conserved.

excess of rare segregating sites per nucleotide (see Table 1). This pattern extends upstream of the coding region, but the 1.4 kb downstream of the exon conforms more to the usual pattern of similar numbers of rare and common polymorphisms. Another intriguing feature associated with exon 2 is a 79-bp element 350 bp upstream of the translation start site. The element consists of two sets of CN dinucleotide repeats arranged as $(\mathrm{CN})_{15}X_{21}$ $(\mathrm{CN})_{13}$ and is conserved in *D. pseudoobscura* although with reduced spacing between the sets of repeats. Ten of the sites are different in *D. simulans*, 9 of which change the N nucleotide in one of the 28 repeats. Of three segregating sites in *D. melanogaster*, only one is common, and it affects a C and is associated with variation for wing shape (Palsson and Gibson 2004, accompanying article). These data strongly suggest that the element is part of a motif that affects *Egfr* transcription in Drosophila. In addition, a complex CAA microsatellite described in the discussion is located 250 bp upstream of the translation start site for exon 2.

The 2.4-kb portion of intron 2 immediately preceding exon 3 is overall the least polymorphic region of the entire *Egfr* locus and includes five stretches of at least 100 bp that are almost monomorphic in *D. melanogaster* and show >50% nucleotide identity in *D. pseudoobscura*. Most of the 102 segregating sites in this region are rare (only 26 common sites observed compared to 57 expected), and similar to exon 2, there is little haplotype structure. By contrast with the sequences flanking exons 1 and 2, which contain 22 common deletions of which half are >4 bp, only 2 long deletions present in just three alleles are found upstream of exon 3. Similarly, only two short deletions are seen relative to the *D. simulans* allele. Although the *cis*-acting regulatory regions have not been mapped in the *Egfr* locus, the phylogenetic and intraspecific shadowing data in Figure 2 strongly suggest intron 2 as a candidate regulatory region.

Exons 3–6 encode the bulk of the DER protein, are highly conserved between species, and show low levels of polymorphism in *D. melanogaster*. Only 4 of the 1325 amino acids in this region differ in *D. simulans*, and there are only two segregating common replacement polymorphisms. The three introns of 170, 66, and 74 bp in the 4.3 kb of sequence covering these four exons have dramatically elevated divergence relative to the exons in both of the other species, as seen in Figure 2, B and C, although no increase in intraspecific polymorphism is observed in the two smaller introns (Figure 2A). Despite these unusual features, there is no formal support for departure from neutrality, perhaps because the intron sequences are so short. Intron 3 is particularly polymorphic, with 1 in every 6 nucleotides harboring a polymorphism. The two ends of the intron, 10 bases from their respective splice sites, show different recombination histories. A remarkable 3-base stretch before the start of exon 4 has polymorphisms with the rare allele at frequencies of 0.20, 0.34, and 0.50 that are almost in linkage equilibrium with one another, whereas four substitutions at a frequency of 0.04 in a 7-base stretch following exon 3 are in perfect linkage disequilibrium. Sequence conservation extends for at least 1 kb 3′ of the termination codon, in both the 3′-untranslated region (UTR) and the intergenic region.

Linkage disequilibrium: Regional variation in the extent of haplotype block structure along the locus is

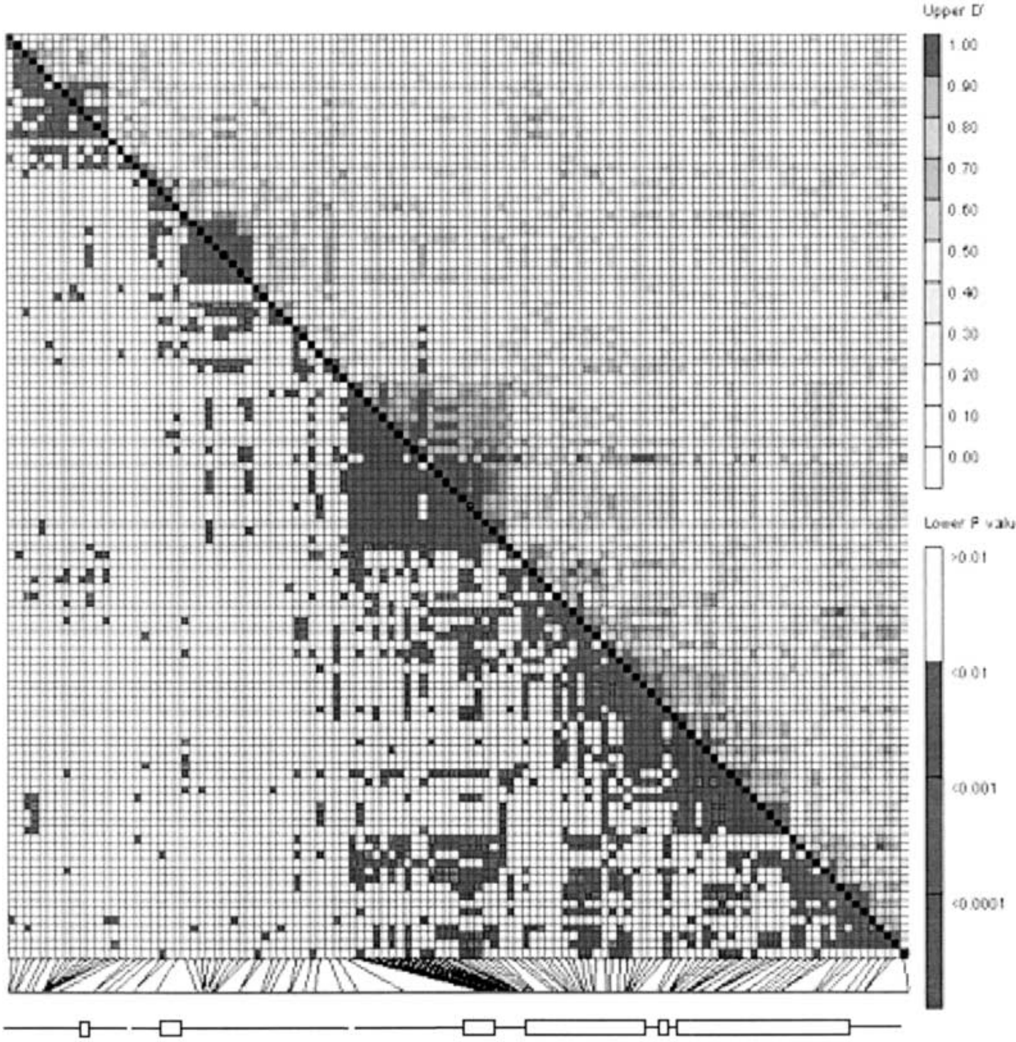

FIGURE 3.—Linkage disequilibrium plot. The extent (D') and significance (P-value) of LD is plotted for each pairwise comparison of 114 common SNPs (more rare allele frequency >0.25, sampled in at least 50 alleles) from 5′ to 3′ (top to bottom and left to right). See supplementary information at http://statgen.ncsu.edu/ggibson/SupplInfo/SupplInfo6.htm for color version from larger set of 238 SNPs with rare allele >0.05. Approximate location of exons and variation in SNP density is indicated below the plot. Three broad regions of LD are evident, particularly in the significance plot, corresponding to the three sequence contigs, but there is considerably reduced LD around exon 2. The large block of LD in the center of the plot corresponds to a highly polymorphic 400-bp region in exon 3 and intron 3.

shown in Figure 3, a plot of the degree (above the diagonal, D') and significance (below the diagonal) of LD between sites along the gene. The plot is for the combined NC and CA data consisting of partial sequence from 210 lines and for clarity includes only SNPs where the less common allele has a frequency of at least 0.25. Two large blocks of strong LD correspond to 1.25 kb in and around exon 1 and the highly polymorphic 400-bp stretch encompassing exon 3 and intron 3. Both the exon 2 and intron 2 sequence stretches are essentially devoid of significant LD, while the 3′ half of the gene including the long exons shows a string of imperfect LD blocks up to 1 kb in length and a suggestion of occasional long-range associations. Significance is a function of the allele frequencies and correlation between them (r^2) as well as the total sample size, whereas D' shows the extent of LD relative to the maximum possible association. Even within the blocks of highly significant LD, adjacent sites are rarely in complete LD, and sites within a few hundred bases of one another can be in linkage disequilibrium. The depth of sampling enabled a comparison of LD between the North American populations, and it proved quite similar as the Pearson correlation between the two populations was 0.86 for r^2 and 0.54 for D'. Table 2 shows that the power to detect pairwise associations that are significant experiment wide rises with sample size, but also depends critically on the absolute count of rare alleles, as argued by LEWONTIN (1995).

Linkage disequilibrium is generally expected to decay with distance, as a result of recombination breaking up haplotypes that are generated as new alleles arise or are introduced by admixture. The plot of LD with distance in Figure 4 confirms that this tends to be the case, with only sites within 50 bp of one another in linkage disequilibrium. LD rarely rises above the background level for sites that are separated by >1 kb. Note that the gaps in the plot without points arise because there were two unsequenced regions of 23.5 and 3 kb in introns 1 and 2. Some long-range associations are expected to arise by chance in the 20,100 pairs of alleles, so most of the r^2 measures beyond 5 kb are probably sampling artifacts. Over a short range, LD can be maintained by factors such as low recombination, gene con-

TABLE 2

Effect of sample size on LD metrics and their significance

	Jackknifed subsamples[a]			Observed		
	$N = 64, \overline{X} \pm$ (SD)	$N = 128, \overline{X} \pm$ (SD)	$N = 255, \overline{X} \pm$ (SD)	A	B	C
Tests	5797.4 (1949.4)	5649.1 (622.7)	5662.5 (29.4)	5671.0	5460.0	13530
Average r^2	0.068 (0.006)	0.062 (0.004)	0.058 (0)	0.058	0.056	0.040
Average D'	0.388 (0.096)	0.394 (0.039)	0.400 (0.001)	0.400	0.252	0.464
>0.05	76.4% (0.026%)	62.2% (0.024%)	47.3% (0.002%)	47.3%	61.4%	63.2%
<0.05	23.6% (0.026%)	37.8% (0.024%)	52.7% (0.002%)	52.7%	38.6%	36.8%
<0.01	13.3% (0.017%)	24.5% (0.019%)	39.2% (0.002%)	39.4%	27.7%	24.8%
<0.001	7.4% (0.01%)	15.3% (0.014%)	27.8% (0.002%)	27.9%	21.0%	16.3%
<0.0001	4.9% (0.009%)	10.7% (0.009%)	20.5% (0.001%)	20.6%	16.8%	11.8%
<0.00001	3.5% (0.008%)	8.2% (0.007%)	16.1% (0.001%)	16.2%	14.1%	9.0%
<0.000001*	2.4% (0.006%)	6.6% (0.005%)	13.0% (0.001%)	13.1%	12.4%	7.0%
<0.0000001	1.7% (0.005%)	5.4% (0.005%)	10.5% (0.001%)	10.5%	10.8%	5.5%

*The Bonferroni significance threshold is 0.000001 for 5000 tests.

[a] Based on 25 random subsamples from the full allele matrix. Sites were extracted on the basis of the minimum number of polymorphic alleles for a site and the absolute count of the less common allele. For $N = 64$, only sites represented by 50 or more alleles were included, if and only if the rare variant was present in 5 or more alleles. These numbers were 100 and 10 for $N = 128$, or 200 and 20 for $N = 255$. The observed sets were: A, 200, 20; B, 150, 20; and C, 200, 5.

version, and epistatic selection, but a comprehensive analysis of these tendencies is beyond this study.

Another way to represent haplotype structure is to plot the actual alleles and haplotype networks, as is done for five representative sequence segments in Figure 5. On the left, each row represents a SNP, while alleles are columns ~0.5 mm wide, and light shading indicates the rare genotype. In the networks on the right, circle diameters are proportional to the number of alleles of a common haplotype, and these are joined by lines with lengths proportional to the number of substitutions distinguishing the haplotypes. The short segments from exons 1 and 5 are examples of true haplotype blocks, with just a handful of alleles in each case having arisen by recombination. Exon 5 provides an example of a dimorphic haplotype, with one common type that is four or five substitutions distinct from the other two common types. The 5′ end of intron 2 by contrast shows several haplotypes and recombination over a stretch of 250 bp. Haplotype dimorphism is also observed in short stretches of exon 3 and intron 3, with deep branches of at least six substitutions separating the two major types in each case, but recombination and/or gene conversion has also generated networks of haplotypes that cannot be generated solely by stepwise mutation. Both of these blocks also present instances of linkage equilibrium between adjacent sites that are interspersed with other sites in LD.

Population structure: Population differentiation was investigated using Wright's estimator of the proportion of variance in allele frequencies attributable to two or more subpopulations relative to the total population, F_{ST} (WEIR and HILL 2002). Figure 6 shows F_{ST} as a sliding window of 10 polymorphic sites along the locus for the two North American populations below the abscissa, and each of these is compared to the smaller Kenyan sample above it. Bars below the plot highlight regions of significance for the three contrasts with four bars representing $P < 0.0001$ and one bar $0.01 < P < 0.05$. For the comparison of North Carolina and California, one window shows a very large F_{ST} that is significant after Bonferroni correction for multiple comparisons ($P <$ 0.0002), but the others must be regarded as marginally significant at best. Analysis of single sites summarized in Table 3 indicates that in all but one of the cases the population structure is focused on a single SNP around which linkage disequilibrium decays rapidly, the exception being the trio of sites 40428, 40458, and 40464 in exon 6.

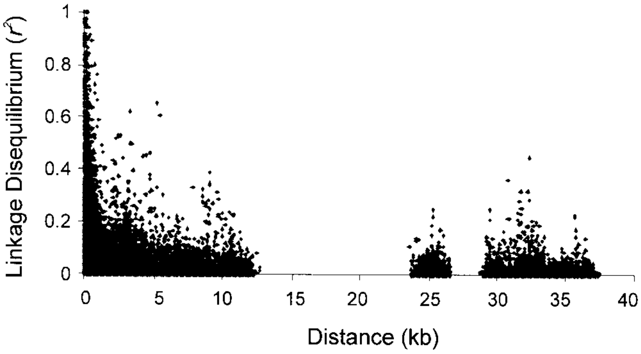

FIGURE 4.—Rate of decay of LD with distance along the locus. Each spot shows the squared correlation coefficient (r^2) between a pair of sites separated by the indicated distance. Gaps arise as a result of the two large gaps of 23.5 and 3.0 kb between contigs. The 20,100 comparisons are derived from 201 alleles with more rare allele frequency >0.1.

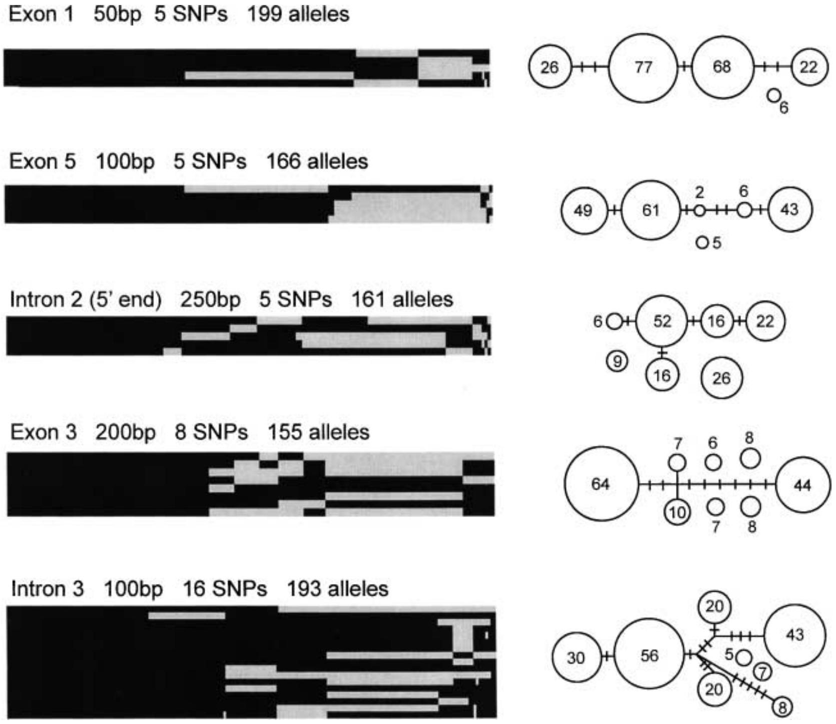
Exon 1 50bp 5 SNPs 199 alleles
26
77
68
22
6
Exon 5 100bp 5 SNPs 166 alleles
49
61
2
6
43
5
Intron 2 (5' end) 250bp 5 SNPs 161 alleles
6
52
16
22
9
16
26
Exon 3 200bp 8 SNPs 155 alleles
7
6
8
64
44
10
7
8
Intron 3 100bp 16 SNPs 193 alleles
20
43
30
56
5
7
20
8

FIGURE 5.—Five representative haplotype blocks. Blocks on the left show alignment of the indicated number of SNPs in the indicated stretch of DNA: solid, the common allele; shaded, the less common allele. Numbers of alleles differ due to incomplete sequence coverage. Networks to the right were drawn by hand to represent the number of alleles of each haplotype class in the sample, and the number of SNPs that separate them (small bars crossing lines joining haplotypes). Haplotypes that must have arisen by recombination or recurrent mutation are not connected with the remainder of the network.

Divergence is much more pronounced between the North American and Kenyan samples, with a maximum F_{ST} of 0.4. There is a tendency for elevated divergence associated with exon 2, intron 2, and exon 6, and it is perhaps noteworthy that the latter two regions are among the least polymorphic in the sample. Population structure can also be seen in the comparison of private allele frequencies, namely alleles that are observed in only one of the populations (SLATKIN 1985). The Kenyan sample has 92 private alleles, almost three times as many as in the Californian sample, which is two and a half times larger. The North American samples share 50 sites that are not observed in Kenya, and 14 of these are common polymorphisms in the sense that the less common allele has a frequency >0.05. Moreover, the Californian sample shares 14 sites with the Kenyans while the larger North Carolina sample shares only 4 sites with the Kenyans.

DISCUSSION

Our survey is 20 times larger than most molecular evolutionary studies, which focus on 2 or 3 kb of 30 or so alleles. Increasing the sample size to >200 alleles has little effect on point estimation of most parameters of nucleotide variation, so for most molecular evolutionary studies it is not necessary. For this reason, PLUZHNIKOV and DONELLY (1996) concluded that sequencing effort should be invested in length, not number, of alleles. However, consistent with SIMONSEN *et al.*'s (1995) simulation study of the power of Tajima's D to detect selective sweeps, we find that large samples decrease the variance

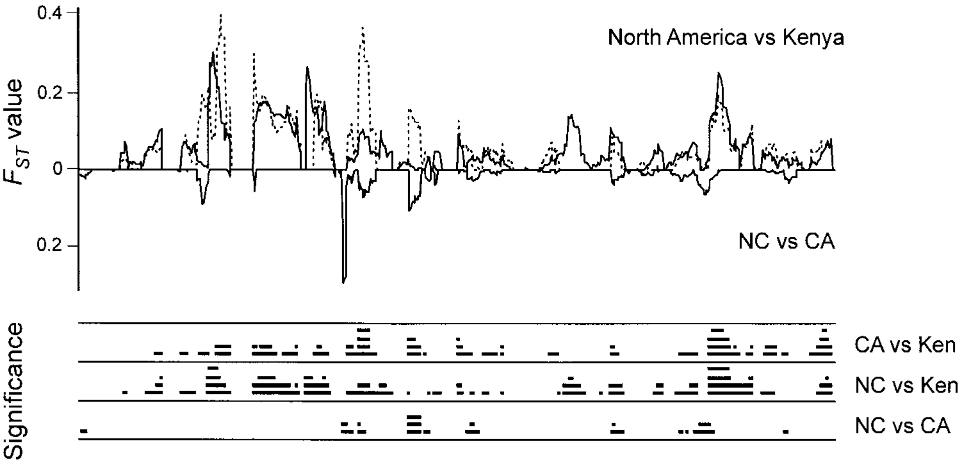

FIGURE 6.—Population differentiation at the *Egfr*. (Top) Plots of F_{ST} for sliding window of 10 SNPs relative to position along the locus for NC *vs.* Kenya (dotted line) and CA *vs.* Kenya (solid line above abscissa) and NC *vs.* CA (below abscissa). (Bottom) Significance of F_{ST} contrasts shown in columns for respective contrasts (one bar, $P < 0.05$; two bars, $P < 0.01$; three bars, $P < 0.001$; four bars, $P < 0.0001$).

TABLE 3

Frequencies and F_{ST} parameters for a comparison of the North Carolina and California populations

		N			NC to CA contrast[b]			Haplotypes[c] (NC)		Haplotypes[c] (CA)	
Site	Variant	NC[a]	CA[a]	Kenya[a]	Window	F_{ST}	*P*-value	*N*	Diversity	*N*	Diversity
35345	A	73	72	NA	1	0.10093	0.0013	2	0.2949	2	0.0526
	C	16	2	NA	10	0.28514	0.0026	4	0.6570	4	0.5833
35697	T	64	10	25	1	0.17558	0.0004	2	0.4807	2	0.3945
	C	43	27	3	10	0.07428	0.0183	6	0.5754	3	0.4917
36214	G	93	43	30	1	0.13641	0.0003	2	0.2817	2	0.4868
	A	19	31	3	10	0.0765	0.0001	10	0.4976	5	0.5889
39010	C	79	67	30	1	0.08844	0.0009	2	0.4652	2	0.2722
	T	46	13	5	10	0.03822	0.0077	8	0.7818	12	0.8129
40428[d]	G	73	31	20	1	0.07552	0.0032	2	0.4806	2	0.4747
	A	49	49	16	10	0.06561	0.0026	6	0.5875	5	0.5309
40464[d]	T	54	51	20	1	0.0688	0.0041	2	0.4917	2	0.4622
	C	70	29	16	10	0.06561	0.0026	6	0.5875	5	0.5309
42023	A	90	36	23	1	0.0958	0.0016	2	0.4175	2	0.4986
	G	38	40	12	10	0.02461	0.0181	10	0.8110	12	0.7625

[a] Absolute counts of the SNP states at the seven sites in the three populations (NA, site not surveyed in the Kenyan sample).

[b] F_{ST} between NC and CA, for individual sites and haplotypes spanning 10 segregating sites with corresponding *P*-value for each estimate.

[c] Diversity and number (*N*) of haplotypes (alleles) per population.

[d] Sites 40428 and 40464 are in nearly complete LD ($D' = 1$, $r^2 = 0.9$, $P < 0.0001$ by Fisher's exact test). They are separated by 38 bp and contribute to the same 10-site haplotype.

of estimates to a degree that is sufficient to markedly increase the power of tests of neutrality. We argue here that detailed analysis of large samples of select regions of the Drosophila genome will provide extra insight into the evolution of population structure, haplotype structure, and complex insertion-deletion polymorphisms.

The only tests of neutrality that produce nominally significant deviations from the null hypothesis are Fu and Li's statistics, of which *D** is plotted in Figure 7 for 100-nucleotide windows along the *Egfr* locus. These sites exceed significant thresholds only for the complete sample: jackknife samples of <100 alleles summarized in supplementary Table 3 (http://statgen.ncsu.edu/ggibson/SupplInfo/SupplInfo6.htm) do not provide any suggestion of departure from neutrality. Of the 341 tests, 21 are significant at $P < 0.05$, but it should also be noted that none of these exceed the conservative Bonferroni threshold for multiple comparisons. Consequently, marginal evidence for purifying selection in the 3′ half of the gene is rendered insignificant by inclusion of the 5′ regions in the analysis. The fact that there is a large excess of negative test statistics might also imply an excess of rare alleles, but Bustamante *et al.*'s (2002) comparison of multiple loci from several species implies that the baseline for comparison is species specific and that negative test statistics are common in Drosophila. What has not been studied carefully is the variance of tests like Tajima's *D* and Fu and Li's *D** within loci and the effect of weak selection on that

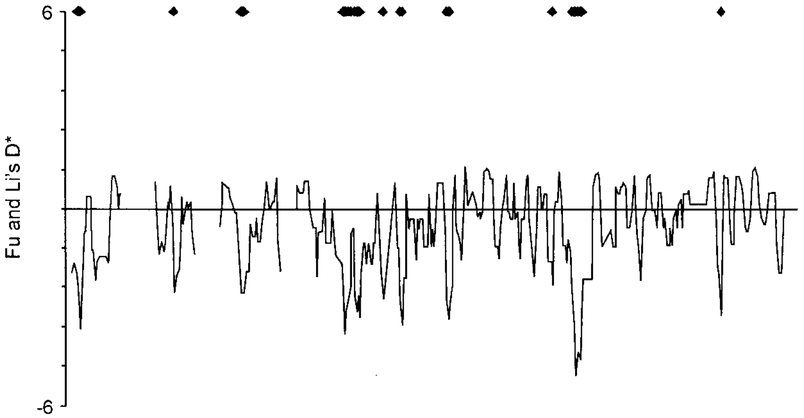

Figure 7.—Weak evidence for purifying selection at the *Egfr* locus. Plot of Fu and Li's *D** statistic (Fu and Li 1993) in 100-bp sliding windows at 25-bp steps along the locus showing generally negative values, consistent with a deficit of intermediate frequency variants due to purifying selection. Of the 341 windows, 21 are significant at $0.05 > P > 0.01$, and 9 at $0.01 > P > 0.001$, and in all cases these are associated with negative values (indicated by solid diamonds). The sample size ranges from 78 in the 3′ UTR to 162 in the exon 1 promoter.

variance. Since none of the *Egfr* regions are even nominally significant in samples of <40 alleles, genome-scale sequencing of this sample size will not have the resolution to address all molecular evolutionary processes.

Haplotype structure: Estimation of haplotype structure is essential both for understanding the evolutionary history of the locus and for determining the density of SNP sampling that will support genotype-phenotype association studies. The two analytical approaches used to infer haplotypes are assessment of the level of linkage disequilibrium (Figures 3 and 4) and phylogenetic analysis of regions of low recombination (Figure 5). Both analyses imply that haplotypes in *Egfr* are generally restricted to fewer than five common polymorphisms that all lie within one 500-bp stretch of sequence. Beyond 1 kb, linkage disequilibrium drops below 10% of maximum, and almost all sites are effectively in linkage equilibrium. The few examples of stronger association are due to either relatively rare alleles or chance, given the very large number of tests (Pritchard and Przeworski 2001). Assessment of the significance of LD is one area where larger sample sizes have a dramatic impact, as the number of significant associations increases >10-fold by measurement of 128 rather than 68 alleles. Comparison of LD in multiple populations is emerging as a crucial aspect of the human HapMap project, and our results confirm the importance of large samples. The documented levels of LD are consistent with other surveys of variation in *D. melanogaster* (Schaeffer and Miller 1993; Richter *et al.* 1997; Langley *et al.* 2000; Zurovcova and Ayala 2002) and imply that haplotypes in the fruitfly are almost two orders of magnitude shorter than those observed in humans (Reich *et al.* 2001; Ardlie *et al.* 2002). Both the low level of linkage disequilibrium and high polymorphism can be attributed to the fact that flies are outcrossing and have had an enormous effective population size for millions of years (Aquadro *et al.* 2001).

More detailed analysis of the haplotype blocks, however, suggests a fine scale to the evolutionary forces acting on the locus. Several authors have remarked that there is in fact an excess of linkage disequilibrium in non-African flies relative to theoretical expectations based on empirical measurement of the recombination rate and neutral mutation parameter (Schaeffer and Miller 1993; Kirby and Stephan 1996; Andolfatto and Przeworski 2000; Wall 2001). Possible biological explanations include pervasive balancing selection or small selective sweeps, a history of admixture in the species, and unequal gene conversion and/or recombination rates along a chromosome (Nachman 2002; Wall *et al.* 2002; Andolfatto and Wall 2003). Figure 5 provides a hint of the latter, since nonoverlapping 250-bp segments of *Egfr* with similar SNP densities clearly produce haplotype networks with distinct topologies ranging from two deep branches to a network of recombined alleles. Similar to our previous observation of bimodality in random sequence stretches (Teeter *et al.* 2000), it appears that several of the putative haplotype blocks in *Egfr* are bimodal. Coalescent theory suggests that the deep branches are not unexpected in a sample of genes, but that their depth and prevalence will be affected by a variety of demographic factors (Slatkin and Hudson 1991). While we do not present new analytical statistics, we hope that the suggestions of variation in haplotype structure across the locus will inspire more theoretical work on the variance of branch lengths under uneven recombination and with weak selection.

An even more subtle pattern in the sequence variation is the existence of strong linkage disequilibrium between sites within 1 kb of one another that are nevertheless separated by several common SNPs in linkage equilibrium. A level of heterogeneity in LD with distance is anticipated due to stochastic evolutionary forces like mutation, genetic drift, and gene conversion but could also be a consequence of population history or directional or epistatic selection. Our analysis of a hypermorphic *Egfr* phenotype in the eye (Dworkin *et al.* 2003) identified one example of significant epistatic interaction between sites 40119 and 40620 in exon 6. Those sites are in LD but are separated by several common and uncoupled SNPs. In this case, linkage disequilibrium reduces the frequency of the hyperactive two-site haplotype, consistent with the possibility that purifying selection contributes to maintenance of the association between the sites. In general, for this mechanism to operate, the selection pressure must be greater than the recombination rate (Lewontin 1964), which is of the order of 10^{-5} events/generation between sites separated by 100 bp in Drosophila euchromatin. Consequently, weak epistatic selection would not be expected to maintain linkage disequilibrium between pairs of sites over much more than a few hundred base pairs. Several instances of departure from monotonic decrease within the linkage disequilibrium blocks are evident in Figure 5, but much greater sampling depth will be required to ascertain whether this is a general feature of the fly genome and to determine the causes of such events.

Population structure: *D. melanogaster* is a human commensal species that is thought to have spread from Africa in the past 15,000 years (Powell 1996) and is conventionally regarded, largely on the basis of allozyme data, as panmictic throughout its range. Numerous recent studies have challenged the latter assumption: several populations from southern Africa show incipient reproductive isolation from the remainder of the species (Wu *et al.* 1995; Ting *et al.* 2001), microsatellite frequencies show that European samples are even more genetically restricted than North American ones (Caracristi and Schlötterer 2003), and examples of clines of genetic variation have been documented (Berry and Kreitman 1993; Verelli and Eanes 2001; Freydenberg *et al.* 2003). The two most common measures of population structure are differences in nucleotide diversity and F_{ST}

TABLE 4

Effects of sampling depth on significance of goodness-of-fit test for indel distributions

	All indels			Deletions			Insertions		
	$N = 20$	$N = 50$	$N = 100$	$N = 20$	$N = 50$	$N = 100$	$N = 20$	$N = 50$	$N = 100$
$\overline{X}$	42.35	49.00	54.81	16.40	19.95	23.08	16.25	18.87	21.27
(SD)	(2.38)	(2.48)	(2.43)	(1.67)	(1.62)	(1.46)	(1.32)	(1.36)	(1.36)
Tests									
>0.05	2.3%	0.8%	0.0%	3.1%	0.5%	0.0%	2.3%	0.8%	0.0%
<0.05	97.7%	99.2%	100.0%	96.9%	99.5%	100.0%	*64.0%*	*60.8%*	*78.4%*
<0.01	82.6%	90.0%	94.9%	84.5%	98.8%	100.0%	22.1%	43.6%	70.4%
<0.001	35.7%	50.4%	72.7%	78.0%	96.5%	99.8%	15.4%	10.2%	6.7%
<0.0001	*13.2%*	*28.8%*	*35.9%*	68.4%	88.5%	93.1%	1.3%	0.0%	0.0%
<0.00001	5.1%	3.8%	1.0%	47.3%	66.7%	84.3%	0.0%	0.0%	0.0%
<0.000001*	0.2%	0.0%	0.0%	29.0%	52.1%	71.4%	0.0%	0.0%	0.0%
<0.0000001	0.0%	0.0%	0.0%	*16.9%*	*41.1%*	*59.9%*	0.0%	0.0%	0.0%

The number of indels segregating and fraction of significant goodness-of-fit tests from 1000 jackknifing samples of sizes 20, 50, and 100 are shown. The observed number of indels, deletions, and insertions in noncoding regions were 63, 27, and 25, respectively (11 could not be classified as deletions or insertions). Italic numbers correspond to the observed significance level for the whole North American data set (220 alleles) for the respective mutation classes.

statistics. Averaged across the entire 10.9 kb of sequence, the Kenyan sample is more diverse than either the Californian or the North Carolinian samples, due to both rare and common SNPs. Among alleles shared among all of the populations, Figure 6 provides strong evidence for allele frequency differences in several regions of the gene between African and American samples. There is no reason to suppose that these differences are not simply due to genetic drift. A more surprising finding is that 50 SNPs not seen in Africa are shared by the CA and NC samples, which implies an extensive period of isolation of the New World flies from their source population prior to their spread across North America. Our data both confirm that a southern African sample is more diverse than derived North American ones (Schlötterer and Harr 2002) and add another layer of evidence for population differentiation in the New World. Glinka *et al.* (2003) observed a similar differentiation between European and African flies in a broad survey of loci on the X chromosome and discussed the possible role of selection in the New World in producing the observed differentiation.

Population stratification can cause false-positive associations between nucleotide variants and phenotypes in population-based linkage disequilibrium mapping, and for this reason it is critical to establish whether there is global evidence for structure in this data set. The depth of sampling in the two North American data sets allows more accurate assessment of population differentiation along the locus. One important question is whether there is structure within a population. Our NC population was collected on a single afternoon from a series of peach bins. Since no two alleles are identical it is unlikely that a single male has contributed in a biased manner to the sample, but it is quite possible that demes congregate at collection points. We tested a random set of 15 SNPs from throughout the locus and found them to be evenly distributed between NC and CA using the Structure algorithm of Pritchard *et al.* (2000). However, Figure 6 also provides suggestive evidence for several SNPs having significantly different allele frequencies on either side of the continent. Further analysis in Table 3 shows that in all but one case the differentiation is restricted to a single site rather than a haplotype. None of these allele frequency differences is even nominally significant for comparisons of <30 alleles from each population (supplementary Table 5 at http://statgen.ncsu.edu/ggibson/SupplInfo/SupplInfo6.htm). It is unclear what could cause a change in frequency of at least 40% for a single site, without affecting closely linked sites, but if selection contributes, it must be very weak relative to the recombination rate, and it is possible that the high F_{ST} values are a sampling artifact.

Even so, if every locus throughout the fly genome harbors just a couple of SNPs that show allele frequency variation, there would be a total of at least 20,000 SNPs differentiating the geographic races, and a combined measure would provide an unambiguous geographic identifier. In this sense, flies are not so different from humans: well over 95% of the diversity is unstructured between geographic regions, but just a couple of percent of the sites may provide some differentiation at the DNA level (Aquadro *et al.* 2001; Rosenberg *et al.* 2002).

Insertion and deletion polymorphism: As insertion and deletion polymorphism constitute only ~10% of segregating polymorphisms in *D. melanogaster*, tests of deviation from neutral models are inevitably less powerful for this class of mutations. Schaeffer (2002) intro-

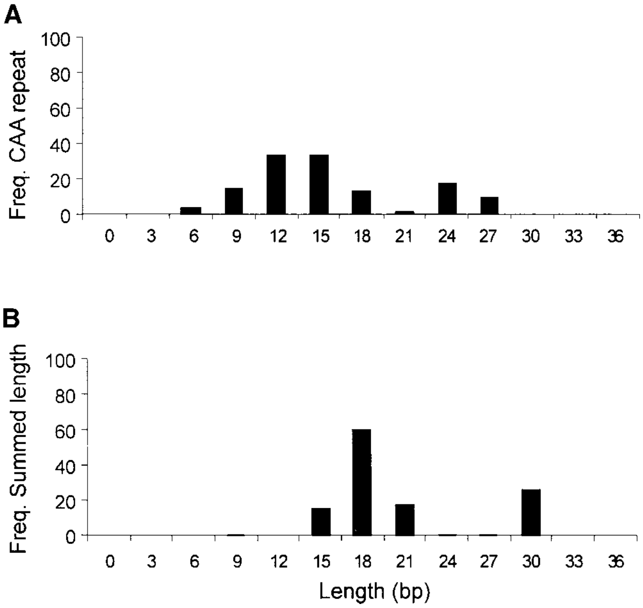

Figure 8.—Nonneutral distribution of microsatellite lengths. (A) Distribution of the most polymorphic trinucleotide repeat ($N = 123$ alleles). (B) The summed length over all three repeats. Note the distillation of the bimodal pattern when considering the length of the whole microsatellite, and that the two peaks at 18 and 30 bp are separated by 12 bp. Length classes 9, 24, and 27 are represented by only a single allele. The invariable CA in the middle of the element is not included in the summed length.

duced a goodness-of-fit test for equal distribution of indels across segments of the *Adh* and *Adh-related* loci in *D. pseudoobscura.* The North American sample has 65 indel variants, of which more than half are <0.05 in frequency, with some evident constraint on indel size: 78% are <10 bp, compared with 56% anticipated by the mutation spectra for Drosophila (Petrov and Hartl 1998). The deletion events are not scattered randomly among the noncoding regions as determined by a goodness-of-fit test ($P < 0.0001$), while the insertions are marginally significant ($P = 0.03$). In particular, there is a deficit of large indels in 1900 bp upstream of exon 3 and in the 3′ UTR whereas larger indels reach high frequencies in the vicinity of exons 1 and 2 and in the three short introns (see supplementary Figure 3 at http://statgen.ncsu.edu/ggibson/SupplInfo/SupplInfo6.htm). Resampling demonstrates that this test does depend quite markedly on sample size, as shown in Table 4.

An even more compelling indication of purifying selection was observed for a complex microsatellite in promoter 2. The consensus sequence $(CAA)_{2\text{-}9}CA(GCA)_{1\text{-}3}(ACC)_{0\text{-}1}$ is flanked by two 35-bp stretches that are nearly invariant in *D. melanogaster* and differ by only six SNPs when compared to *D. pseudoobscura.* The most variable repeat has a tendency toward bimodal distribution of length classes that is greatly enhanced after addition of the shorter trinucleotides (Figure 8 and supplementary Table 4 at http://statgen.ncsu.edu/ggibson/SupplInfo/SupplInfo6.htm). The distribution of summed repeat lengths deviates significantly from normal (Kolmogorov-Smirnov goodness-of-fit test, $P < 0.01$), being extremely platykurtic (kurtosis −0.41, $P < 0.0001$) and right skewed (0.91, $P < 0.0001$) as assessed by 1000 jackknife iterations. The fact that the two classes are separated by 12 bp, which corresponds to a turn in the helix, suggests the hypothesis that the length distribution may be molded by selective constraints on the orientation of the conserved flanking regions. These results should encourage a wider survey of length distributions of composite microsatellites and their relation to the functional dissection of evolutionarily conserved noncoding regions.

Evolution of *Egfr*: The DER protein is among the least variable of all fly proteins. Disregarding alternate exon 1, there are just a handful of rare, and two common, replacement polymorphisms in >1400 amino acids of sequence. Three of the proteins that initiate the signaling cascade downstream of DER, namely DRK, RAS, and DSOR, are actually even less polymorphic and are identical in a sibling species (Gasperini and Gibson 1999; Riley *et al.* 2003), but DER is clearly subject to strong purifying selection. Aside from structural constraints, two possible explanations for functional constraint are high levels of pleiotropy (Waxman and Peck 1998) and occupancy of a critical control point in a biochemical pathway (Nijhout 2002; Olsen *et al.* 2002; Riley *et al.* 2003). *Egfr* exhibits both of these features, being involved in dozens of developmental processes and integrating intercellular signaling events. Given Fraser *et al.*'s (2002) demonstration of a relationship between sequence constraint and the complexity of protein-protein interactions in yeast, it may also be informative to conduct comparative phylogenetic analyses of the other seven receptor tyrosine kinase loci in the fly genome.

The high level of amino acid polymorphism in exon 1, including a derived second initiator codon, is quite surprising given that alternate exon 2 is just as constrained as the common portion of the protein. Most of the variation is in the presumptive signal peptide, so it does not alter the mature protein. The two exons are both used in most tissues, but exon 1 is also used in a subset of adult neurons (Lev *et al.* 1985; Schejter *et al.* 1986). Lesokhin *et al.* (1999) tested the effect of two of the exon 1 polymorphisms that happened to be present in an $Egfr^{Ellipse}$ gain-of-function allele (W15R and L21W) and found them to be qualitatively neutral with respect to phosphorylation of DER targets in transgenic flies. Since sequence conservation is also much reduced in *D. pseudoobscura* and undetectable in the mosquito *Anopheles gambiae,* presumably precise function of exon 1 is less important to the organism than that of exon 2.

As remarked, the current sample size decreases the variance in estimators of linkage disequilibrium, molecular evolution, and population subdivision and may also influence analysis of the length distribution of indels and composite microsatellites. We regard the most striking aspect of the analysis, however, to be the subtle shifts in sequence diversity along the gene. Statistical measures associated with individual blocks of 1 or 2 kb differ only marginally if at all from expectations of neutral theory, but it is the diversity of patterns within such a small region of the genome that calls for explanation. As for the vast majority of the genome, there is nothing particularly noteworthy about diversity in the *Egfr*. Yet there are subtle signs that various modes of selection, drift within populations, uneven recombination rates, and migration have all helped shape the variation in this typical Drosophila gene.

A number of people were instrumental in helping with the sequence analysis, including Bryon Sosinski and Regina Brierley at the North Carolina State University Genome Research Laboratory and April Duty and Naruo Nikoh in our lab. We also thank Roland Carrillo for help with inbreeding the North Carolina lines, Sergey Nuzhdin for the Californian lines, Naruo Nikoh for help with analysis, and Ed Buckler for use of his Tassel software. A.P. was funded in part by awards from the American Scandinavian Foundation and North Atlantic Treaty Organization, and the project was funded by the National Institutes of Health (R01 GM61600).

LITERATURE CITED

Andolfatto, P., and M. Przeworski, 2000 A genome-wide departure from the standard neutral model in natural populations of Drosophila. Genetics **156:** 257–268.

Andolfatto, P., and J. Wall, 2003 Linkage disequilibrium patterns across a recombination gradient in African *Drosophila melanogaster*. Genetics **165:** 1289–1305.

Aquadro, C. F., D. Bauer and F. Reed, 2001 Genome-wide variation in the human and fruitfly: a comparison. Curr. Opin. Genet. Dev. **11:** 627–634.

Ardlie, K. G., L. Kruglyak and M. Seielstad, 2002 Patterns of linkage disequilibrium in the human genome. Nat. Rev. Genet. **3:** 299–309.

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.

Bustamante, C. D., R. Nielsen, S. Sawyer, K. Olsen, M. D. Purugganan *et al.*, 2002 The cost of inbreeding in Arabidopsis. Nature **416:** 531–534.

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

Clifford, R., and T. Schüpbach, 1994 Molecular analysis of the Drosophila EGF receptor homolog reveals that several genetically defined classes of alleles cluster in subdomains of the receptor protein. Genetics **137:** 531–550.

Duchek, P., and P. Rorth, 2001 Guidance of cell migration by EGF receptor signaling during Drosophila oogenesis. Science **291:** 131–133.

Dworkin, I. M., A. Palsson, K. Birdsall and G. Gibson, 2003 Evidence that *Egfr* contributes to cryptic genetic variation for photoreceptor determination in natural populations of *Drosophila melanogaster*. Curr. Biol. **13:** 1888–1893.

Eanes, W. F., 1999 Analysis of selection on enzyme polymorphisms. Annu. Rev. Ecol. Syst. **30:** 301–326.

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

Fraser, H. B., A. Hirsh, L. Steinmetz, C. Scharfe and M. W. Feldman, 2002 Evolutionary rate in the protein interaction network. Science **296:** 750–752.

Freeman, M., 1998 Complexity of EGF receptor signaling revealed in Drosophila. Curr. Opin. Genet. Dev. **8:** 407–411.

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

Fu, Y. X., and W. H. Li, 1993 Statistical tests of neutrality of mutations. Genetics **133:** 693–709.

Galindo, M. I., S. Bishop, S. Greig and J. P. Couso, 2002 Leg patterning driven by proximal-distal interactions and EGFR signaling. Science **297:** 256–259.

Gasperini, R., and G. Gibson, 1999 Absence of protein polymorphism in the Ras genes of *Drosophila melanogaster*. J. Mol. Evol. **49:** 583–590.

Gellon, G., and W. McGinnis, 1998 Shaping animal body plans in development and evolution by modulation of Hox expression patterns. BioEssays **20:** 116–125.

Glinka, S., L. Ometto, S. Mousset, W. Stephan and D. De Lorenzo, 2003 Demography and natural selection have shaped genetic variation in *Drosophila melanogaster*: a multi-locus approach. Genetics **165:** 1269–1278.

Hudson, R. R., M. Kreitman and M. Aguadé, 1987 A test of neutral molecular evolution based on nucleotide data. Genetics **116:** 153–159.

Johnson-Hamlet, M., and L. A. Perkins, 2002 Analysis of corkscrew signaling in the Drosophila epidermal growth factor receptor pathway during myogenesis. Genetics **159:** 1073–1087.

Kirby, D. A., and W. Stephan, 1996 Multi-locus selection and the structure of variation at the *white* gene of *Drosophila melanogaster*. Genetics **144:** 635–645.

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

Lesokhin, A. M., S. Y. Yu, J. Katz and N. E. Baker, 1999 Several levels of EGF receptor signaling during photoreceptor specification in wild-type, Ellipse and null mutant Drosophila. Dev. Biol. **205:** 129–144.

Lev, Z., B. Z. Shilo and Z. Kimchie, 1985 Developmental changes in expression of the *Drosophila melanogaster* epidermal growth factor receptor gene. Dev. Biol. **110:** 499–502.

Lewontin, R. C., 1964 The interaction of selection and linkage. I. General considerations; heterotic models. Genetics **49:** 49–67.

Lewontin, R. C., 1988 On measures of gametic disequilibrium. Genetics **120:** 849–852.

Lewontin, R. C., 1995 The detection of linkage disequilibrium in molecular sequence data. Genetics **140:** 377–388.

Mayor, C., M. Brudno, J. Schwartz, A. Poliakov, E. M. Rubin *et al.*, 2000 Vista: visualizing global DNA sequence alignments of arbitrary length. Bioinformatics **16:** 1046–1047.

McDonald, J. H., and M. Kreitman, 1991 Adaptive protein evolution at the *Adh* locus in Drosophila. Nature **351:** 652–654.

Moriyama, E. N., and J. R. Powell, 1996 Intraspecific nuclear DNA variation in Drosophila. Mol. Biol. Evol. **13:** 261–277.

Nachman, M. W., 2002 Variation in recombination rate across the genome: evidence and implications. Curr. Opin. Genet. Dev. **12:** 657–663.

Nicholas, K. B., H. Nicholas and D. Deerfield, 1997 GeneDoc: analysis and visualization of genetic variation. EMBO News **4:** 14.

Nijhout, H. F., 2002 The nature of robustness in development. BioEssays **24:** 553–563.

Olsen, K. M., A. Womack, A. Garrett, J. Suddith and M. D. Purugganan, 2002 Contrasting evolutionary forces in the *Arabidopsis thaliana* floral developmental pathway. Genetics **160:** 1641–1650.

Palsson, A., 2003 Molecular quantitative genetics of wing shape in Drosophila melanogaster. Ph.D. Thesis, North Carolina State University, Raleigh, NC.

Palsson, A., and G. Gibson, 2004 Association between nucleotide variation in *Egfr* and wing shape in *Drosophila melanogaster*. Genetics **167:** 1187–1198.

Petrov, D. A., and D. L. Hartl, 1998 High rate of DNA loss in the *D. melanogaster* and *D. virilis* species groups. Mol. Biol. Evol. **15:** 293–302.

Pluzhnikov, A., and P. Donnelly, 1996 Optimal sequencing strategy for surveying molecular genetic diversity. Genetics **144:** 1247–1262.

Powell, J. R., 1996 *Progress and Prospects in Evolutionary Biology: The Drosophila Model.* Oxford University Press, New York.

Pritchard, J. K., and M. Przeworski, 2001 Linkage disequilibrium in humans: models and data. Am. J. Hum. Genet. **69:** 1–14.

Pritchard, J. K., M. Stephens and P. Donnelly, 2000 Inference of population structure using multilocus genotype data. Genetics **155:** 945–959.

Reich, D. E., M. Cargill, S. Bolk, J. Ireland, P. Sabeti *et al.*, 2001 Linkage disequilibrium in the human genome. Nature **411:** 199–204.

Richter, B., M. Long, R. C. Lewontin and E. Nitasaka, 1997 Nucleotide variation and conservation at the *dpp* locus, a gene controlling early development in Drosophila. Genetics **145:** 311–323.

Riley, R., W. Jin and G. Gibson, 2003 Contrasting selection pressures on components of the Ras-mediated signal transduction pathway in Drosophila. Mol. Ecol. **12:** 1315–1323.

Rosenberg, N. A., J. K. Pritchard, J. Weber, H. M. Cann, K. K. Kidd *et al.*, 2002 Genetic structure of human populations. Science **298:** 2381–2385.

Rozas, J., and R. Rozas, 1999 DnaSP version 3: an integrated program for molecular population genetics and molecular evolution analysis. Bioinformatics **15:** 174–175.

Schaeffer, S. W., 2002 Molecular population genetics of sequence length diversity in the *Adh* region of *Drosophila pseudoobscura.* Genet. Res. **80:** 163–175.

Schaeffer, S. W., and E. L. Miller, 1993 Estimates of linkage disequilibrium and the recombination parameter determined from segregating nucleotide sites in the alcohol dehydrogenase region of *Drosophila pseudoobscura.* Genetics **135:** 541–552.

Schejter, E. D., D. Segal, L. Glazer and B. Z. Shilo, 1986 Alternative 5′ exons and tissue-specific expression of the Drosophila EGF receptor homolog transcripts. Cell **46:** 1091–1101.

Schlötterer, C., and B. Harr, 2002 Single nucleotide polymorphisms derived from ancestral populations show no evidence for biased diversity estimates in *Drosophila melanogaster.* Mol. Ecol. **11:** 947–950.

Schneider, S., D. Roessli and L. Excoffier, 2000 *Arlequin: A Software for Population Genetics Data Analysis*, Version 2.0. Genetics and Biometry Lab, University of Geneva, Geneva.

Shilo, B. Z., 2003 Signaling by the Drosophila epidermal growth factor receptor pathway during development. Exp. Cell Res. **284:** 140–149.

Simonsen, K. L., G. A. Churchill and C. F. Aquadro, 1995 Properties of statistical tests of neutrality for DNA polymorphism data. Genetics **141:** 413–429.

Slatkin, M., 1985 Rare alleles as indicators of gene flow. Evolution **39:** 53–65.

Slatkin, M., and R. R. Hudson, 1991 Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. Genetics **129:** 555–562.

Stein, R. A., and J. V. Staros, 2000 Evolutionary analysis of the ErbB receptor and ligand families. J. Mol. Evol. **50:** 397–412.

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

Teeter, K., M. Naeemuddin, R. Gasperini, E. Zimmerman, K. P. White *et al.*, 2000 Haplotype dimorphism in a SNP collection from *Drosophila melanogaster.* J. Exp. Zool. **288:** 63–75.

Thompson, J. D., D. G. Higgins and T. J. Gibson, 1994 CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. Nucleic Acids Res. **22:** 4673–4680.

Ting, C. T., A. Takahashi and C.-I Wu, 2001 Incipient speciation by sexual isolation in Drosophila: concurrent evolution at multiple loci. Proc. Natl. Acad. Sci. USA **98:** 6709–6713.

Upton, G. J. G., 1982 A comparison of alternative tests for the 2 × 2 comparative trial. J. R. Stat. Soc. A **145:** 86–105.

Verelli, B. C., and W. F. Eanes, 2001 Clinal variation for amino acid polymorphisms at the *Pgm* locus in *Drosophila melanogaster.* Genetics **157:** 1649–1663.

Wall, J. D., 2001 Insights from linked single nucleotide polymorphisms: what we can learn from linkage disequilibrium. Curr. Opin. Genet. Dev. **11:** 647–651.

Wall, J. D., P. Andolfatto and M. Przeworski, 2002 Testing models of selection and demography in *Drosophila simulans.* Genetics **162:** 203–216.

Wang, S. H., A. Simcox and G. Campbell, 2000 Dual role for Drosophila epidermal growth factor receptor signaling in early wing disc development. Genes Dev. **14:** 2271–2276.

Watterson, G. A., 1975 On the number of segregating sites in genetical models without recombination. Theor. Popul. Biol. **7:** 256–276.

Waxman, D., and J. R. Peck, 1998 Pleiotropy and the preservation of perfection. Science **279:** 1210–1213.

Weir, B. S., 1996 *Genetic Data Analysis II.* Sinauer Associates, Sunderland, MA.

Weir, B. S., and W. G. Hill, 2002 Estimating F-statistics. Annu. Rev. Genet. **36:** 721–750.

Wu, C.-I, H. Hollocher, D. J. Begun, C. F. Aquadro, Y. Xu *et al.*, 1995 Sexual isolation in *Drosophila melanogaster*: a possible case of incipient speciation. Proc. Natl. Acad. Sci. USA **92:** 2519–2523.

Yang, L., and N. E. Baker, 2003 Cell cycle withdrawal, progression, and cell survival regulation by EGFR and its effectors in the differentiating Drosophila eye. Dev. Cell **4:** 359–369.

Yang, H. P., and S. V. Nuzhdin, 2003 Fitness costs of Doc expression are insufficient to stabilize its copy number in *Drosophila melanogaster.* Mol. Biol. Evol. **20:** 800–804.

Zecca, M., and G. Struhl, 2002 Control of growth and patterning of the Drosophila wing imaginal disc by EGFR-mediated signaling. Development **129:** 1369–1376.

Zurovcova, M., and F. J. Ayala, 2002 Polymorphism patterns in two tightly linked developmental genes, *Idgf1* and *Idgf3*, of *Drosophila melanogaster.* Genetics **162:** 177–188.

Communicating editor: M. A. F. Noor