JOURNAL OF EXPERIMENTAL ZOOLOGY (MOL DEV EVOL) 288:63–75 (2000)

Haplotype Dimorphism in a SNP Collection From *Drosophila melanogaster*

KATHERINE TEETER,[1] MOHAMMED NAEEMUDDIN,[2]
ROBERT GASPERINI,[1] ERIKA ZIMMERMAN,[3] KEVIN P. WHITE,[4]
ROGER HOSKINS,[2] AND GREG GIBSON[1,3]*
[1]*Department of Biology, University of Michigan, Ann Arbor, Michigan 48109*
[2]*Berkeley Drosophila Genome Project, Lawrence Berkeley National Laboratory, Berkeley, California 94720*
[3]*Department of Genetics, North Carolina State University, Raleigh, North Carolina 27695*
[4]*Department of Biochemistry, Stanford DNA Sequencing and Technology Center, Palo Alto, California 94306*

ABSTRACT A moderate resolution single nucleotide polymorphism (SNP) map of the genome of *Drosophila melanogaster* that is designed for use in quantitative genetic mapping is described. Seventeen approximately 500 nucleotide gene sequences spaced at 10 to 20 centimorgan intervals were combined with 49 shorter sequence tag sites (STSs) at 5 to 10 centimorgan intervals to generate a map that should not leave any gaps greater than one half of a chromosome arm when any two wild type lines are compared. Of 20 markers with sufficient polymorphism to construct haplotype cladograms, 13 showed evidence for two divergent classes of haplotype. The possible mechanisms for and implications of the unexpected finding that two thirds of all short gene sequences in *D. melanogaster* may be dimorphic are discussed, including the suggestion that admixture between two separate lineages may have been a major event in the history of the species. *J. Exp. Zool. (Mol. Dev. Evol.) 288:63–75, 2000.*

Despite the many advantages of *Drosophila* for genetic analysis, the development of molecular markers for use in quantitative genetic mapping has lagged behind that of other model organisms. In the past few years, several QTL detection studies have been published, each using an independently derived map based on PCR-RFLPs (Liu et al., '96), retrotransposon insertion sites (Long et al., '95; Nuzhdin et al., '97; Gurganus et al., '99), or microsatellites (Gibson et al., '99). Although each of these methods has their advantages, they do not yield readily to high-throughput scoring methodologies. For this reason, there is a movement in genome-wide mapping studies toward the use of single nucleotide polymorphisms (SNPs: Wang et al., '98), which are amenable to a variety of high-throughput techniques for the detection of these polymorphisms. Here we report the construction of a moderate resolution SNP map of the *D. melanogaster* genome with markers at approximately five centimorgan intervals that should be appropriate for QTL mapping starting with essentially any pair of wild type inbred lines.

The three major advantages of a SNP map are resolution, ease of genotyping, and versatility. Since an average of approximately one in every 200 nucleotides differ between any two chromosomes of *D. melanogaster* (Moriyama and Powell, '96), SNPs are detectable at a very high density simply by sequence comparison. Furthermore, any stretch of one kilobase of sequence should contain several polymorphisms and hence present high haplotype diversity, facilitating comparison of many different lines using the same marker. Most SNPs can be detected without the need for costly and time-consuming gel electrophoresis. High-throughput detection techniques include denaturing high pressure liquid chromatography (DHPLC: Underhill et al., '97), allele-specific oligonucleotide (ASO) hybridization on nylon membranes (Saiki et al., '86), a ligation-mediated fluorescence method (Chen et al., '98), MALDI-TOF (Ross et al., '98), and DNA array hybridiza-

Grant sponsor: David and Lucille Packard Foundation; Grant number: 96-5154A

K. Teeter and M. Naeemuddin contributed equally to this work.

*Correspondence to: Greg Gibson, Department of Genetics, Gardner Hall, North Carolina State University, Raleigh, NC 27695-7614. E-mail: ggibson@unity.ncsu.edu

Received 14 August 1999; Accepted 13 December 1999

tion on glass slides by oligonucleotide tiling (Chee et al, '96) or with resequencing protocols that are under development (Pastinen et al., '97). These technologies will soon enable the high throughput and cost-effective genotyping of hundreds of markers in several thousand individuals.

SNP markers are codominant if scored for both alternate alleles and hence can be used in a variety of experimental designs where heterozygotes and both homozygote classes need to be distinguished. Each marker utilizes a pair of PCR primers that is common to all strains (and in most cases will also amplify DNA from the sibling species) and has a known map position, facilitating comparison of different studies. In most cases, once marker-trait associations have been detected with one pair of lines, the same markers—sometimes using alternate linked SNPs—will be suitable for following the most significant QTLs in different backgrounds. Importantly, high resolution mapping and allelic association studies between nucleotide polymorphisms in candidate genes and trait variation also require SNP detection, so an SNP map will provide a means for transitions between the various phases of quantitative genetic studies (Lynch and Walsh, '97).

Using two strategies, the results presented below confirm that DNA sequencing is an efficient way to identify SNPs in *Drosophila*. In the first strategy, 20 gene fragments were selected according to map position using sequence information obtained from GenBank. Seventeen of these 500 base pair fragments proved to be sufficiently polymorphic to support distinction of the vast majority of a set of ten randomly sampled alleles. In parallel, a set of over 750 short sequence tag sites (STSs) had been sequenced between two common lab stocks for use in mapping of Mendelian mutations (R.H. and M.N., unpublished data). We subsequently sequenced a subset of 49 of these markers in six other inbred lines to expand the SNP collection, provide data as to which sites may show high heterozygosity in wild type lines, and to fill in some gaps in the map.

While analyzing these data, an unexpected pattern of haplotype dimorphism was observed in about two thirds of the markers that were sufficiently polymorphic to attempt to construct cladograms. We recently reported a similar phenomenon in one of the Ras genes of *Drosophila*, *Dras2* (Gasperini and Gibson, '99), and several studies of allozyme polymorphisms have remarked on the divergence of haplotype classes adjacent to sites that are thought to experience balancing selection (e.g., Kreitman and Hudson, '91). However, haplotype dimorphism on a genome-wide basis has not been described. This may be because over the several kilobases that are typically examined, recombination is sufficient to break up linkage disequilibrium so that over the full length of any gene the polymorphism is effectively jumbled. Fortuitously, the marker sequences considered here are short enough that linkage disequilibrium appears to be retained in small samples. While the pattern of nucleotide diversity for any one gene is not statistically different from standard neutral equilibrium models, the collection of over a dozen markers with evidence for dimorphism is intriguing enough to suggest a need for much greater population genetic sampling throughout the genome.

MATERIALS AND METHODS

Fly stocks

To generate nearly isogenic lines, wild type strains of *D. melanogaster* were inbred by pairwise sib-mating for 10 to 50 generations. Canton S and Oregon R were obtained from Dr. Gerald Rubin, and 1st, 2nd, and 3rd chromosomes were isogenized by passage over balancer chromosomes. Samarkand, inbred for over 260 generations, was obtained from Dr. Trudy Mackay, and the remaining stocks were generated by G.G. Lines W6 through W29 derived from a collection of near-isofemales collected by B. Wallace from various localities around the world (see Gibson and van Helden '97, Table 1, for a complete list; W6 is from Capetown South Africa, W11 from the Pyrenees, Spain, and W28 and W29 from Kenya). Lines A6 through A20 were derived from isofemales trapped in an Ann Arbor, Michigan, fruit market in summer 1997.

Sequencing

PCR products were amplified from genomic DNA prepared from one to several flies of each inbred line. For the sequencing of gene fragments, each PCR product was purified from an agarose gel and directly sequenced using one of the amplification primers. ABI dye-terminator chemistry was used with the ABI 377 automated slab gel sequencing system to obtain reads of over 500 nucleotides in most cases. For the STS sequencing, PCR products were treated with exonuclease I and shrimp alkaline phosphatase and sequenced directly without gel purification (Werle et al. '94) using ABI BigDye terminator chemistry and an ABI 377 DNA sequencer. PCR primers for the gene

sequences are listed in Table 1, and for the STSs can be obtained over the internet from the Berkeley *Drosophila* Genome Project (BDGP: http://www.frutifly.org).

Gene sequences were aligned and polymorphisms detected by eye using Sequence Navigator software (ABI Prism). The complete set of STS sequences were analyzed together after aligning them using Phrap, then detecting polymorphisms with the Phred (Ewing et al., '98) and Consed (Gordon et al., '98) programs obtained as described at the following web site of the University of Washington Genome Center: http://www.phrap.org/consed/consed.html#howToGet. Some manual realignment was necessary, and all traces were also scanned by eye to confirm the presence of unambiguous SNPs and rare heterozygotes.

Data analysis

Nucleotide diversity listed in Table 1 was determined using the equation (Tajima, '93):

$$\pi = \Sigma\Sigma\, 2k_{ij}/(n^2 - n)$$

where n is the number of alleles sequenced, k is the number of differences per site between a pair of alleles, and the sum is over all possible pairwise combinations of alleles i and j. It thus provides a measure of the average distance between any two alleles for each marker. The proportion of sequence comparisons that differ between any two alleles is a more useful measure in judging the potential of each marker for SNP detection, and this was calculated as the proportion of all possible pairwise comparisons that differ by at least one polymorphism for the particular marker.

The set of 49 STS sequences in Table 2 were chosen solely on the basis of genomic location from a larger set of 304 markers that were already known to have at least one (and an average of 2.1) SNP difference(s) between Oregon R and Canton S (M. Ellis, R.H. and M.N., unpublished data). There were no differences in SNP frequency between the 49 markers in Table 2 and the remaining 255 markers of the larger set. However, this sample is likely to be slightly more polymorphic than a completely random set of markers because over 60% of the chosen STSs differ between each of the pairs of lines, with a low value of 50% for the two most similar lines (W11 and Oregon R), compared with 40% for Oregon R and Canton S (304 dimorphic STS of a pool of 751 STS sequences compared).

Cladograms in Figure 4 were drawn by hand. Identical sequences and those that differ by one or two sites were grouped, and then samples were removed that appeared to be recombinant (that is, they consisted of the left end of one haplotype and the right end of another) or showed evidence for double recombination. Automated algorithms are not necessary with such small samples, and those that are available do not adjust for the presence of recombination. Diagrams can be derived for any marker using the data in Figures 1 and 2.

ASO analysis

Males from each of 75 isofemale lines, most of which were trapped in Kenya by R.C. Woodruff (lines 3676–3750 of the old Mid-American Stock Center, now maintained in G.G.'s lab), were crossed to W6, and a single progeny was chosen for DNA preparation. Using PCR products blotted onto a nylon filter (HybondN+; Pharmacia Biotech) under vacuum, genotypes were inferred from the presence or absence of hybridization of a ^{32}P-labeled 15 mer allele-specific oligonucleotide (ASO: Saiki et al., '86) that was designed to be complementary to the non-W6 SNP allele . Two replicate blots with individuals in different arrays were performed for each gene (*tkv* and *fz*). The following wash temperatures and ASO oligonucleotides (polymorphic sites underlined) were used in a hybridization buffer containing 5× SSPE and 0.1% SDS:

tkv241	TTCATGTTGATGTAC (40°)
tkv303	ATATGTAGCATACTT (25°)
tkv356	AGTTAAGACGCCCAT (32°)
tkv392	GCCTCAAAATAGGAC (25°)
tkv407	GTCCTCATGTGTTTT (30°)
fz43	TGGCCGTTTGCTACT (52°)
fz85	GACTGGGGGCGGGCG (60°)
fz375	ATTCGTTCACAGCAT (45°)
fz449	CGAGTTTTCTTCTTTT (25°)

For *fz449*, two different hybridization intensities were observed, possibly indicating recombination between the three adjacent polymorphisms (C/T, T/C, and TT/—) in three of the 19 individuals that hybridized to the probe. Haplotypes and linkage can be inferred directly with this design since the ASOs do not hybridize to the W6 reference allele.

RESULTS

Construction of a low resolution SNP map

As a first step toward construction of an SNP map *of Drosophila melanogaster*, we sequenced ap-

TABLE 1. Genes and nucleotide diversity parameters

Gene	Location	Primers	N	Poly	Length	Prop	1000π	Type of sequence
Achaete	1B	GGCTGAGAGGAACAACTGATAC TTTCAGTGTGCTAACTTTGCTC	11	8 (1)	450	0.91	4.30	5′ end (half coding)
Swallow	5E	GAGGACGACTATGATGAGGATG TCGCCTTGAATAGAAACCAAC	11	4	500	0.87	2.76	exon 2 coding + introns
Sevenless	10A	CAGGAGGATCTGTTTCTGGAC CGGAGAGTAGAGGACTTCGTC	8	9 (2)	450	0.79	5.48	coding
Folded gastrulation	20B	GGAGACTATGACTACGGCGAC TGAGGAGCTTGAATTAGCAAC	11	7 (1)	550	0.88	4.73	coding
Nina A	21E	GTCTGCAGAGGGTATGCCATC TGGGTTTCATTCCATTCGGAG	4	9	500	0.83	9.00	coding + introns
Thickveins	25D	GTTGAATGCGCAGTTCCGACC TGCTCTTACAGGCTAGTCATC	4	13	500	0.83	14.30	3′ non-coding
Numb	30B	AGCAATTGTCGCCAGACTTGC GGTGTACCGCTACACTGACAC	4	4 (1)	500	1.00	4.67	3′ non-coding
Apterous	41F	CTTGGACTAACGGATGCTCAG GCTTGGTAAAAACATTGCCAGC	3	1	400	0.67	1.67	3′ non-coding
Even skipped	46C	ACTGCATAACAATGGAACCCG ATGGCTGCCATGACTTTCGG	4	4 (1)	500	0.83	3.33	5′ regulatory
Transformer 2	51B	GTGTGCAATATAGCAGGGAATC TTCGTTCGCGATCGCGTAGATC	4	3 (2)	500	1.00	1.00	mainly introns
Diptericin	56A	TGCAGTTCACCATTGCCGTCG CAATTTGGCCATTCTTAGCTGG	4	9 (1)	350	0.83	13.33	coding + 3′ ntr
Seizure	60B	GGATTTGGCAATGTGGCACCG GGTATATACACGGATTCTCGC	2	0	700	0	0.00	introns + exons
Roughened	62B	AGAGATATACGAAGGATATAC GGTTTTTGGAAGTCTTATAGCA	11	3	500	0.71	2.25	coding
Tryptophan hydroxylase	66A	CAGTGGAGAAACCCGAGAATC CCTCGACTATGTAAGCCGAATC	11	5 (1)	450	0.47	3.11	5′ non-coding
Frizzled	70D	GTGCTCACCTTCTTGATTGAC GGATTTCCACAGAACTTACCTTC	11	10 (3)	500	0.85	5.05	exons + introns
Dras1	85D	CTACCGTGAGCAGATCAAGCG CTCGCAGCCTTTCAAACGACAC	12	5	800	0.87	2.30	coding + intron
Dopamine receptor	88B	CCGTATCACGTATCCGACCAC GCAAGTGACATGTGTCACTCC	4	3	500	0.67	4.00	3′ end (half coding)
14-3-3-epsilon	90F	GTCCTCTTTACAGAAATGGTGG GGTTCCATGTTGTTGTTCTTGA	11	1	550	0.55	0.99	introns + exons
Dromyosuppressin	96A	ATGGTTGCCGGCCACTGATCA GTTGCACATAGGACACGTTCC	11	10(1)	350	0.91	7.17	5′ end
Tailless	100B	CCTCACAGCAGACAACACAAC GGCATTCTCGGACTCGTAGAC	12	6 (1)	290	0.80	8.57	coding + intron

TABLE 2. Polymorphic EST markers[1]

Marker	Loc.	Length	N	Poly	Haps	Prop.	BiA?[2]
Dm2977	1C	150	7	3	3	0.52	–
Dm1729	2B	200	7	2	3	0.52	–
Dm3238	4C	170	8	1	2	0.57	–
Dm2931	6B	150	7	1	2	0.48	–
Dm0426	8C	200	8	2	4	0.79	–
Dm3169	8D	200	8	4	4	0.64	–
Dm3746	10F	240	8	5	5	0.86	Y
Dm0478	13A	170	8	2	2	0.43	+
Dm3790	15B	190	7	2	3	0.67	+
Dm0501	17A	180	7	4	3	0.67	Y
Dm0505	19E	180	8	6	5	0.52	N
Dm0447	21D	150	7	5	3	0.76	–
Dm0605	23E	170	8	1	2	0.57	–
Dm0342	27C	170	8	4	5	0.86	–
Dm2818	30A	160	8	2	3	0.61	–
Dm4048	32A	170	7	5	5	0.81	Y
Dm3982	34B	240	5	2	2	0.60	–
Dm2390	36D	180	8	3	2	0.43	Y
Dm0877	38B	200	8	3	4	0.75	+
Dm0976	41C	200	8	2	3	0.61	+
Dm0746	44A	170	8	1	2	0.43	–
Dm0750	46B	150	8	1	2	0.54	–
Dm0084	48C	170	8	3	4	0.75	–
Dm2233	50C	150	8	2	3	0.64	–
Dm0064	50D	170	7	4	4	0.81	–
Dm2633	51E	220	8	1	2	0.25	–
Dm2192	54A	170	8	3	4	0.75	–
Dm0800	56B	140	6	4	3	0.67	–
Dm0932	58B	180	7	2	4	0.81	N
Dm0891	60C	155	6	2	3	0.73	–
Dm0688	61D	140	7	2	3	0.67	+
Dm2349	63D	160	7	3	4	0.86	N
Dm0925	66A	170	7	7	6	0.95	Y
Dm3873	67F	200	8	6	4	0.75	Y
Dm3681	70A	160	7	3	4	0.71	–
Dm0184	71B	150	8	1	2	0.25	–
Dm2140	74A	170	7	2	2	0.48	+
Dm0970	76B	180	8	1	2	0.57	–
Dm0980	77B	180	8	1	2	0.54	–
Dm2723	80B	230	8	1	2	0.25	–
Dm2182	83C	220	8	1	2	0.43	–
Dm2758	86C	200	8	2	3	0.61	–
Dm0038	88B	240	8	3	4	0.75	–
Dm2494	90A	170	6	3	3	0.60	+
Dm3710	92B	130	6	2	3	0.73	+
Dm2730	94C	220	8	10	6	0.89	Y
Dm1624	96B	180	5	4	3	0.70	–
Dm2759	98A	150	8	2	3	0.68	–
Dm2288	100B	190	6	2	3	0.73	–

[1]N, number of sequences; Poly, number of polymorphisms per marker; Haps, number of distinct haplotypes; Prop., proportion of pairwise comparisons that differ for at least one site; BiA?, indicates whether the marker may be bi-allelic. See http://www4.ncsu.edu/~ggibson/flySNPs for an expanded list of SNP sequences.

[2]Y, Yes; N, no; +, maybe; –, insufficient polymorphism to judge.

proximately 500 nucleotides of 20 genes in four highly inbred wild-type lines. Subsequently, up to ten more wild-type alleles were sequenced for genes on the first and third chromosomes. The genes were chosen to provide three or four evenly spaced markers per chromosome arm. Table 1 shows the cytological location, primers, and type of sequence for each gene, as well as the number of alleles sequenced, number of polymorphisms (including SNPs and small indels), the approxi-

Chromosome 1

	1B *Achaete*	5E *swa*	10A *sevenless*	20B *fog*
		2222	777777777	1111111
	11223345	1555	566777889	2235556
	47130542	9379	933114462	2933788
	82921181	9384	356048384	5155292
A6	AT-GCTGT	TCGG	AC-CAGTC-	.....
A8	GTGGCTGT	TCAG	ACGCAGTC-	.....
A16	.....	TCAT	.....	GAT3CGT
A18	.....	....	.....	.GT-TCC
A19	GTGGCTAT	....	.....	GGT-TCC
W6	ATGGATGT	TCAG	AC-TAGAC-	GGC-CCC
W7	AT-AATGA	TCAT	.....	GAT3CGT
W11	AT-G....	TCAT	AC-CAGTC-	GAT3CGT
W14	GTGGCTGT	TCGG	.....	TAT3CGT
W22	AT-GATGT	TCAG	CC-TAGACT	.....
W23	AT-GATGT	TCGT	AC-CAGTC-	TAT3CGT
W28	GTGGCTGT	ACAG	AC-CAGTC-	GGT3CCC
W29	ACGGCAGT	TAGG	AT-TGATTA	GGT3CCC
	41514111	1144	111311211	2514355
	NNDNVVNV	VVNV	VNINNNVNV	VNNDNVN

Chromosome 3

	62B *R*	66A *Tph*	70D *frizzled*	85D *Dras1*	88B *DopR*	96A *DMS*	100B *tll*
		11111		33333	111		
	13	01123	2245555666	25689	799	122222233	112
	817	90472	3795668444	48931	234	8001235923	589003
	694	02213	7907096012	63753	410	4649044863	485364
A6	...	GCTTI	.....	.....	...	.....	CACCAC
A8	...	AGAAI	TGITT-CTC-	.....	...	.....	CGTCTC
A16	...	.....	CAITT-CCT2	.....	...	GTCCC9GCCG	.....
A18	CGG	.....	TGITT-CTC-	AAGGT	...	TTCAC9GCCG	CGTCTC
A19	...	GCTTI	CAITT-CCT2	.....	...	.....	CACTAT
W6	TAA	GCTTI	CAITT-CCT2	A..CC	TTA	TTCAC9GTCG	TATCAT
W7	CAG	GCTTI	CAITT-CCC2	GAGGT	...	TTCAC9GTCG	TACCAC
W11	CAG	GCTTI	.....	.....	TTA	TTCAC9GTCG	TACCAC
W14	CAG	GCTTI	.....	AAGCC	...	TGCAC9GTCG	CGTCTC
W22	...	GCTTI	CAITT-CCT2	.....	...	.....	TACCAC
W23	CAG	GCTTI	TGITT-CCT2	AGTCC	...	.....	TACCAC
W28	CGA	AGAAI	CGITT-CCT2	AAGGC	CCC	TTCCC9ACCG	CACCAC
W29	CAG	GCTA-	CG-AA4TTC-	AAGGC	CCC	GTGCT-GCAA	CGTCTC
	123	22231	3511111343	11132	222	2113111411	545142
	NNN	NVVVD	NNDVVINNND	NNVVN	NNV	VVVVNDNNVN	NNNNVN

Chromosome 2

	21E *nina A*	25D *thickveins*	30B *numb*	46C *eve*	51B *tra2*	56A *Diptericin*
		2222222333333	2222			222222222
	11233344	7888999000001	3468	3444	577	334445556
	306501223	0001289478992	8767	6236	044	397892280
	398565160	0896684172254	4360	3971	026	820501469
W6	GACTGGGTG	TGTACAATCATAT	T-TG	CTT6	T44	CCGACGCAT
W11	GACGAGGTG	TGTACAATCATAT	GITG	CTT6	C44	CCGACGCAT
W28	GACTGGGTA	TTGTTATAAGCGA	GICA	CTC6	C--	CAACTGCAA
W29	TCTTGCACA	GTGATGTACACGA	TICG	ACC-	C4-	AAACTAA-T
	111111112	1221212211222	2121	1121	112	122221111
	VVNVNVNNN	VVVVNNVVVNNNV	VDNN	VNND	NDD	VVNVNNVDV

Fig. 1. SNP polymorphisms in 17 randomly distributed gene segments. The nucleotide positions relative to GenBank accession sequences are indicated vertically, and lines run horizontally (abbreviated on the left). Numbers at the bottom of each column indicate the number of alleles with the more rare genotype. A dot (.) indicates sequence not obtained or readable; - indicates deletion (of the single base shown, or the number of bases shown for the alternate allele); N, transition; V, transversion. Number and letter above each gene indicates cytological location. GenBank accession numbers are: *achaete,* M17120; *swa,* X56023; *sev,* J03158; *fog,* U03717; *ninaA,* X14769; *tkv,* U1442; *numb,* M27815; *eve,* X78903; *tra2,* X57484; *Dpt,* Z11728; *R,* X02200; *Tph,* X98116; *fz,* X54650; *Dras1*, X73219; *DopR,* X77234; *Tll,* M34639. The DMS sequence is contained within BACR48A04 of the *Drosophila* genome project and was initially supplied by R. Nichols.

mate length of sequence, and an estimate of nucleotide diversity, π. The column labeled "Prop" indicates the proportion of comparisons for which at least one polymorphism distinguishes each pair of sequenced alleles.

A complete list of polymorphisms is provided in Figure 1. One half of the polymorphisms detected (58 of 112) were singletons, and the vast majority were base substitutions, with only 14 indels in the sample. Nucleotide diversity fell within the known range for *Drosophila* (Moriyama and Powell, '96), averaging 0.005 polymorphisms (including indels) per base pair with a maximum value of 0.014 for part of the 3′ non-coding region of the *thickveins* (*tkv*) locus. Three loci (*seizure*, *apterous* and *14-3-3-epsilon*) were monomorphic or nearly so, and hence do not provide SNP polymorphisms. The remaining 17 loci averaged 6.6 polymorphic sites, and typically any two alleles have different haplotypes in over 80% of pairwise comparisons.

Construction of a moderate resolution SNP map

To increase the probability of finding markers at 10 to 15 centimorgan intervals for any comparison of two inbred lines, and to fill in a few gaps in the map, a second SNP-detection strategy was adopted. We sequenced 49 shorter (on average 180 base pairs) sequence tag site (STS) sequences in eight strains including three common inbred laboratory stocks, Canton S, Oregon R, and Samarkand. Nucleotide diversity was similar to that for the longer sequences (average = 0.006), but the total number of polymorphisms per marker was accordingly lower, averaging just 2.8 substitutions or indels. Nevertheless, each marker

Chromosome 1

	1C	2B	4C	6B	8C	8D	10F	13A	15B	17A	19E
	2977	1729	3238	2931	0426	3169	3746	0478	3790	0501	0505
	1	33	2	1	12	222	11111	12	12	2233	122223
	555	14	1	9	34	9045	11567	64	13	7712	934890
	584	73	8	7	56	4430	12996	72	72	5940	999575
W6	TGT	TA	T	D	GA	CCGT	ACTTC	AA	AT	CGGT	ACICTA
W29	TAA	CA	A	.	GT	CTAC	AACTC	GC	..	....	ATICTA
W11	AGT	TA	T	D	AA	CCGT	GCCGC	AA	AT	CGGT	GTICTT
A8	AGT	TA	A	D	GA	ACAT	AACTC	GC	AT	CGGT	ACICTA
A20	...	TA	A	D	GT	CCGT	AACTT	AA	CT	CGGT	GTICTT
Sam	AGT	TG	T	2	GA	CCGT	AACGC	AA	CC	TACC	GTDCAT
C-S	AGT	..	A	D	AT	CCAT	AACTC	AA	AT	TAGC	ATITTA
O-R	AGT	TA	T	2	GT	CCGT	GCCGC	AA	CC	TACC	ACICTA
	211	11	4	2	24	1131	23131	22	32	3323	331113
	VNV	NN	V	I	NV	VNNN	NVNVN	NV	VN	NNVN	NNDNVV

Chromosome 2

	21D	23E	27C	30A	32A	34B	36D	38C	41C	44A	46B	48C	50C	50D	51E	54A	56B	58C	60B
	0447	0605	0342	2818	4048	3982	2390	0877	0976	0746	0750	0084	2233	0064	2633	2192	0800	0932	0891
	11	2	1223	1	11111	22	11	222	12	1	2	1	22	111	1	222	1111	22	12
	56906	1	5361	87	01778	19	934	033	68	6	9	882	66	9138	9	699	7799	49	90
	91338	5	9735	27	45092	08	511	126	81	3	3	682	28	5608	9	123	3689	72	60
W6	DDCGG	G	TCTG	TT	CAGCC	TG	TTT	T6I	C2	5	G	CAG	GG	AAAA	C	GTG	ACAG	..	CG
W29	22TTC	G	TCAG	TA	...CC	TG	ACA	ADI	C2	D	G	CTG	GG	AGTT	C	GCG	ACGT	AA	CG
W11	22TTC	A	TTTG	TT	DAGCC	GT	TTT	A6I	C2	5	G	CAG	TG	AAAT	C	TCG	TAGG	GA	CG
A8	DDCGG	A	TCAG	AT	DTCCC	..	TTT	ADI	C2	D	T	AAG	GG	AAAT	C	GCC	ACAG	GA	GA
A20	2DTTC	H	TCTG	TT	DAGAG	TG	TTT	ADI	TD	5	T	CAG	GC	A..T	C	TCG	TAGG	AC	GG
Sam	22TTC	H	ACTG	TT	DTCCC	..	TTT	ADD	C2	5	G	CAG	GC	AGTT	C	TCG	ACAG	GC	GA
C-S	2DTTC	G	TCTG	AT	DAGCC	..	TTT	ADI	CD	5	T	AAG	GG	TAAT	C	GCC	....	GA	..
O-R	....	A	ACTT	TT	DTCCC	GT	ACA	T6I	TD	5	G	CAA	HG	AGTT	T	TCG	....	GC	..
	24222	4	2121	21	13311	22	222	231	23	2	3	211	22	1331	1	412	2232	23	32
	DDNVV	N	NNVV	VV	IVVVV	VV	VNV	VID	ND	D	V	VVN	VV	VNVV	N	VNV	VVNV	NV	VN

Chromosome 3

	61D	63D	66A	67F	70A	74A	76B	77B	77E	80B	83C	86C	88B	90A	92B	94C	96B	98A	1OOB
	0688	2349	0925	3873	3681	2140	970	0184	980	2723	2182	2758	0038	2494	3710	2730	1624	2759	2288
	11	111	2222233	112222	112	1	2	2	2	1	1	12	111	111	12	1111122222	11	11	1
	38	458	2237734	340146	260	89	2	3	7	1	1	94	458	444	71	0157901245	8925	11	85
	26	040	4743788	895577	059	55	6	7	1	9	6	78	327	789	12	5731243867	0205	49	55
W6	CC	AT1	TGDCCA2	C7GDCG	A1G	AA	A	T	C	T	C	AA	TTA	GT4	TT	TGCATCACAG	....	GG	CA
W29	..	GT1	AGDCCA2	T7G1CA	G1G	AA	C	C	C	A	C	AA	TTA	GT4	CC	TACGTCAAGG	TAAA	GG	CG
W11	AG	GTD	AA2CAT2	T7GDCG	G1G	AA	A	T	A	T	C	GA	TTG	GT4	TC	TGCATCGAGG	AACG	AG	CG
A8	CG	GA1	AA2CAT2	T7GDCG	G1A	AA	A	T	C	T	T	AA	GGA	AGD	TT	TGCATCGAGG	....	GG	..
A20	CC	HA1	.....	CDTDAA	GDG	HH	C	T	C	T	C	AA	TTG	...	CC	TGCGCTAAAA	ATCG	AG	CA
Sam	CC	GA1	AGDCGA2	T7GDCG	G1G	CG	C	T	C	T	C	AA	TTA	GT4	TT	HGCHCCAAGA	ATCG	AG	CA
C-S	CC	...	AA2CAAD	CDTDAA	...	CG	C	T	A	T	T	GA	TAA	ATD	..	CGHACTAAAA	ATCG	AG	TG
O-R	AG	AT1	AA2TAT2	T7GDCG	G1G	AA	A	T	A	T	C	AG	TTA	...	..	TGTGCTAAAA	....	AA	..
	23	331	1331331	322123	111	32	4	1	3	1	2	21	122	212	23	2142432144	1211	31	13
	VV	NVD	VNDNNVD	NDVIVN	NDN	VN	V	N	V	V	N	NN	VVN	NVD	NN	NNNNNNNVNN	VVVN	NN	NN

Fig. 2. SNP polymorphisms in 49 STS segments. Legend as for Figure 1, with gene names replaced by the number of the STS (297 = Dm2977, etc). Nucleotide positions refer to position in the STS sequence.

typically distinguishes a pair of alleles in two thirds of comparisons, as indicated in Table 2. These STS markers may be slightly enriched for polymorphism relative to a completely random set of markers since they were prescreened to differ between Oregon R and Canton S (see Materials and Methods).

A complete list of SNP polymorphisms is provided in Figure 2. The fraction of singletons (44/138 polymorphisms) was slightly lower than for the longer gene sequences, and there were 16 indels. Two sites provided evidence for recurrent mutation in the form of two different nucleotide substitutions. Three markers were unusually polymorphic (Dm-

0925 at 66A, Dm3873 at 67F, and Dm2730 at 94C). The locations of each marker together with a graphical representation of its relative efficiency at distinguishing pairs of alleles is shown in Figure 3. As expected, there is a tendency for increased polymorphism in regions of higher crossover frequency (Begun and Aquadro, '92), but the sequences are too short to perform statistical tests of this relationship (data not shown).

The frequencies and proportions of transitions, transversions, and indels in the combined sample of 250 polymorphisms are shown in Table 3. There were no significant differences between the gene and STS samples. Overall, there were slightly more transitions than transversions, but surprisingly the proportion of transversions was elevated on the second chromosome relative to the first and third chromosomes. The difference is only marginally significant at the 5% level by a chi-square test, and is considerably more pronounced in the set of gene sequences than in the STS data set.

Haplotype dimorphism

Several of the markers in both data sets have sufficient polymorphism to qualitatively assign haplotypes. Two thirds of these markers appear to present two classes of haplotype. Of the ten genes on chromosomes 1 and 3, six produce two clades that are separated by between three and five polymorphisms, as shown in Figure 4. This is most obvious for *Tph*, where eight alleles are identical and differ by four sites from two other alleles. Similarly, *frizzled* (*fz*) has four identical alleles that are five sites divergent from another pair of alleles, while three alleles appear to be recombinants. In most cases, the W29 alleles do not fall into either common haplotype class, and are either putatively recombinant, have several extra unique polymorphisms, or both. For the other four genes, the polymorphism is more evenly spread among the alleles and no obvious haplotype classes are seen.

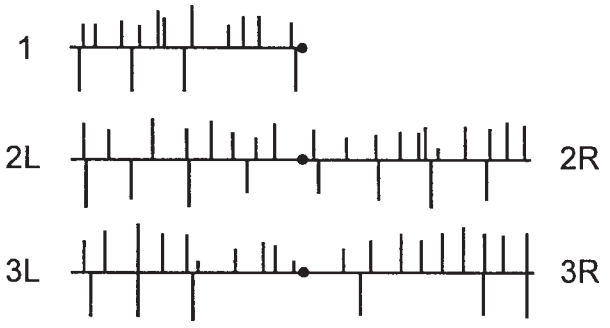

Fig. 3. SNP map showing location and degree of polymorphism of markers. Each horizontal line represents the indicated chromosome arm (1, 2L and 2R, 3L and 3R), with centromeres represented as solid circles. Vertical lines above each chromosome line represent STS markers, with the length of the line proportional to the fraction of pairwise comparisons that are expected to differ by at least one SNP polymorphism at the marker (longer lines correspond to more polymorphic markers). There is a tendency for reduced polymorphism near the telomeres and centromeres, as evidenced by the length of the lines near these regions. Vertical lines below each chromosome line represent the set of gene markers.

TABLE 3. Number of frequencies of types of mutation by chromosome

Type of mutation	Chr. 1	Chr. 2	Chr. 3	Total
Transitions	32 (0.53)	33 (0.35)	52 (0.55)	117 (0.47)
Transversions	23 (0.38)	50 (0.53)	31 (0.33)	104 (0.42)
Indels	5 (0.08)	12 (0.13)	12 (0.13)	29 (0.12)

A similar pattern can be seen in the set of STS sequences, although in this case the majority of markers have too little polymorphism to perform an analysis. The eight most polymorphic markers (Dm3746, Dm0501, Dm0447, Dm4048, Dm2390, Dm0925, Dm3873, and Dm2730) all show some clear evidence for two haplotype classes, while only three markers (Dm0505, Dm0932, and Dm2349) do not appear to be at all dimorphic. Furthermore, of the fifteen remaining markers with just two non-singleton polymorphisms, in ten cases (Dm0478, Dm3790, Dm0877, Dm0976, Dm0064, Dm0891, Dm0688, Dm2140, Dm2194, and Dm3710) the two more rare substitutions are found in the same alleles, consistent with the existence of two haplotype classes.

This observation implies that there should be significant linkage disequilibrium between SNPs within each marker. To confirm that this is the case, we determined the haplotypes of 75 randomly chosen alleles isolated from a collection of Kenyan isofemale lines, over five sites in *tkv*, and four sites in *fz*. Males of each line were crossed to the W6 isogenic stock, for which the haplotype had been determined by sequencing, and one heterozygous offspring was chosen for genotyping using the ASO method. The sites and frequencies of the non-W6 allele in the sample are shown in Figure 5, along with an estimate of D', the ratio of observed to maximum possible linkage disequilibrium, between adjacent pairs of SNPs (Weir, '96). Numbers in brackets show the proportion of haplotypes containing just one non-W6 allele for each pairwise comparison. There is highly significant linkage disequilibrium between three com-

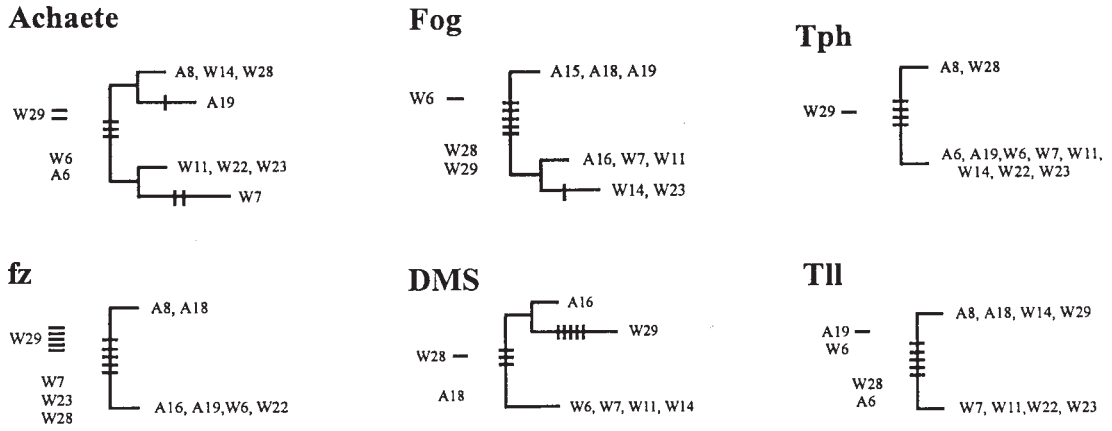

Fig. 4. Hand-drawn cladograms showing haplotype dimorphism in six genes. The short bars crossing each arm of the cladograms represent SNP differences. Alleles with the same haplotype are written adjacent to each branch. Allele names to the left of each figure indicate haplotypes that appear to have arisen by recombination between the two main classes of haplotype, with SNPs unique to these alleles again represented by short horizontal bars. Note that different lines tend to share the same haplotypes for each marker, but that W6, W28 and W29 are generally more divergent or tend to have recombinant haplotypes.

mon polymorphisms separated by 165 nucleotides in *tkv* ($P < 0.001$), and complete linkage disequilibrium between three SNPs separated by 365 nucleotides in *fz*, although there has been considerable random assortment between these sites and the first SNP just 42 nucleotides 5′ in the *fz* sequence.

DISCUSSION

Utility of the SNP map

The aims of this study were to identify a set of evenly distributed SNPs that might be useful for quantitative genetic mapping using wild type lines, and to compare the efficiencies of two strategies for the detection of SNPs. To this end, we opted for greater volume over double-stranded coverage and have analyzed nearly one quarter of a megabase of sequence. It is doubtful that sequencing the reverse strand of each marker would significantly improve SNP detection, but the estimates of population genetic parameters should be regarded with due caution, especially since they are derived from just eight to ten alleles. If there is a bias due to inaccurate scoring, it is likely to be toward underestimation of nucleotide diversity since we have only included polymorphisms that were unambiguous, or in a couple of cases present in two or more individuals. Putative heterozygotes where two nucleotides appeared with similar intensity at a site were not included unless the polymorphism was found in another individual, but such instances were likely to be extremely rare, as observed heterozygosity in the highly inbred lines was less than one percent.

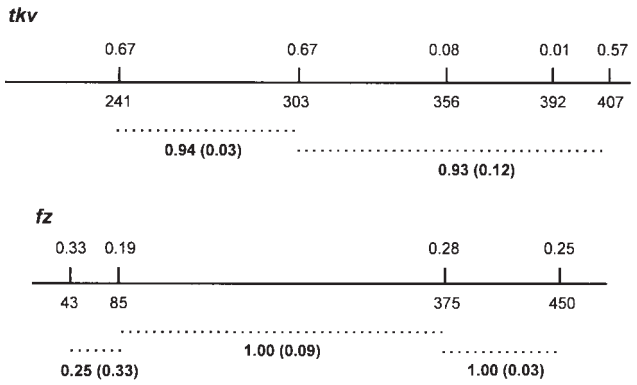

Fig. 5. Linkage disequilibrium in the *thickveins* and *frizzled* genes. The nucleotide positions relative to the 5' sequencing primer are shown below the horizontal line, with the frequency of the non-W6 allele above it. Proportions in bold below dotted lines indicate *D′* (the ratio of observed to maximum possible values of the linkage disequilibrium parameter for the observed allele frequencies), with the fraction of haplotypes for which just one of the two SNPs being compared is the non-W6 form, in brackets.

For the purposes of SNP identification, polymorphism is high enough in *D. melanogaster* to make sequencing of randomly selected short sequences an efficient method, obviating the need for prescreening using techniques such as DHPLC (Underhill et al., '97). The STS sequences average just 180 base pairs in length, but in many cases had two or three polymorphisms. The drawback of using such short sequences is that they are not sufficiently diverse to ensure that a sequence difference will be observed in the majority of comparisons of any two individuals. However, with a

high enough density of markers, this does not present a problem for QTL mapping purposes. Markers every ten centimorgans are sufficient for first-pass QTL mapping (Zeng, '93) and it is generally unlikely that each pair of three adjacent STSs will contain the same sequence in any two individuals. Combining the STS and gene sequences, the largest gap left when any pair of the lines that we have sequenced is just one half of a chromosome arm. It should also be noted that a high density set of SNP markers that distinguish Oregon R and Canton S has also been generated with markers every few hundred kilobases (R.H., M.N. and M. Ellis, in preparation). This data will be a useful starting ground if higher density maps are to be used for QTL mapping in the future.

We have established a web site (http://www4.ncsu.edu/~ggibson/flySNPs) with a list of ASO markers that in principle provides sufficient coverage to enable first pass QTL mapping with progeny derived from any set of parents. Selection of 30 of these SNP sequences should provide at least 20 useful markers without having to sequence the parents, particularly if one of the common lab stocks (Oregon R, Canton S, or Samarkand) is used as one of the parents. However, we caution that there is heterozygosity in a few of the markers. Subsequently, the remaining markers could be used to initiate higher resolution mapping near QTL peaks. A fraction of the polymorphisms also affect restriction sites and so might alternatively be used for PCR-RFLP detection.

A further implication of the data is that sequencing of 500 base pair segments of the genome should be a very efficient way of finding polymorphisms for high resolution mapping once a region of interest has been identified. By designing PCR primers one kilobase apart, the probability that any marker will contain a SNP that distinguishes any two alleles is very high given that automated DNA sequencers routinely provide over 500 base pairs of readable sequence and reactions can be primed from either end. With the availability of the complete sequence of the *Drosophila* genome, there is essentially no limit to the resolution that SNP detection can provide for gene mapping, and this method should be a useful adjunct even where other markers such as microsatellites or AFLPs are used for a low resolution scan. Microarray-based methods for SNP detection are being developed (Chee et al., '96; Pastinen et al., '97) that should make SNPs the quickest and most cost-effective technique for QTL and association mapping in *Drosophila*.

Existence of haplotype dimorphism

One third of the polymorphisms detected by both approaches are "common" in the sense that the more rare variants are present in three or more of the sequenced alleles and hence have frequencies between 0.25 and 0.5. Inspection of Figures 1 and 2 reveals that in those cases where there are two or more of these common alleles in one marker, they tend to be found in the same individuals. This observation led us to construct haplotype cladograms where possible, and thence to the surprising finding that two thirds of the more highly polymorphic loci appear to be represented by two classes of common haplotype as shown in Figure 4. In each case, a set of alleles that differ from one another by one or two sites share between three and six sites that are not shared with most of the remaining alleles. For some of the markers, there are a few alleles that are easily explained as recombinant between the two haplotype classes, which is to be expected since linkage disequilibrium in *Drosophila* is well known to decay over several hundred base pairs in regions of normal recombination (Zapata and Alvarez, '93). In general, the individuals contributing to each haplotype cluster are different for each marker, indicating a history of recombination between markers. There is no obvious differentiation by population since the Ann Arbor and Kenyan haplotypes mix with the other haplotypes.

The statistical significance of this unexpected distribution of haplotype variation is difficult to establish, since there is little power to detect linkage disequilibrium or otherwise to support the inference of two distinct haplotype clusters in samples of just a dozen alleles. Data across markers cannot be pooled to provide a multi-site test of association, because of the recombination between markers. The ASO analysis provides some evidence that sequencing tends to overestimate the extent of non-random associations, both because of the small sample size and the possibility that there is population structure in the data (see Gasperini and Gibson, '99, for another example of loss of evidence for linkage disequilibrium in larger samples obtained by ASO genotyping). Although there was highly significant linkage disequilibrium between all sites within the two genes examined in a sample of 75 alleles, the larger sample highlights that this is against a backdrop of variable SNP frequencies. For example, *tkv* polymorphisms 356 and 392 are only found on chromosomes with the other non-W6 SNPs, but there were only 6 and 1 representatives respec-

tively of each in the large sample, whereas the other non-W6 SNPs are the more common allele. These sites would not then contribute to the divergence between most representatives of the two haplotype classes for this locus. By contrast, *fz* polymorphism 375 was only seen in one of 10 sequences from the worldwide sample, but has a frequency of 0.28 in the Kenyan sample, so may contribute significantly to haplotype divergence. Such observations suggest that supplementing sequence analysis by sampling SNPs for hundreds of alleles will aid in describing the history and distribution of molecular variation in this species (Zapata and Alvarez, '93).

The existence of an excess of within-locus linkage disequilibrium in *D. melanogaster* (and *D. simulans*) has been inferred by Andolfatto and Przeworski (manuscript submitted). They found that estimates of the per nucleotide recombination rate, based on analyses of DNA sequences, are consistently significantly lower than those calculated from direct observation of crossing-over in laboratory crosses. There are several possible explanations for such a result, but the most likely seems to be that non-random association between SNPs decreases the estimate of the rate at which recombination must operate in natural populations. Our data provides direct evidence for such linkage disequilibrium on a relatively unbiased sample of sequences. Our loci were not chosen *a priori* on the basis that they are likely to be polymorphic, or to encode a particular class of gene. In the case of the STSs, which in general derive from the ends of P1 clones that were used to assemble the physical map of the *D. melanogaster* genome, they are only biased to the small extent attributable to preselection of markers that distinguish two common lab stocks.

Implications of haplotype dimorphism

Assuming that the existence of two haplotype classes in more than one half of the sampled sequences from the twelve strains that we considered is representative of a larger population, at least three explanations can be considered: demographic history, selection, and admixture.

The first possibility is that haplotype dimorphism is actually built into the ancestry of any sample of haplotypes. While it seems unlikely that scattering of SNPs at random over a set of haplotypes would frequently lead to two classes of haplotype separated by more polymorphisms than occur within each class, such a random distribution is not actually the correct null model. Slatkin and Hudson ('91) showed that the history of coalescent events imposes a significant correlation structure on the pairwise comparison of alleles, even in populations of constant size and with little demographic structure. For this reason, the existence of deep branches in intraspecific gene trees is not unexpected, but the question is whether the depth of the branches (the number of substitutions separating haplotype classes relative to the diversity within each clade) falls within the range predicted by standard neutral theory. This is impossible to answer with small samples if there is uncertainty in the frequency of each individual SNP, as implied by the ASO analysis. Just as importantly, coalescent histories are affected by parameters such as the ratio of the recombination rate to the neutral mutation rate for each locus, demographic factors (migration, population subdivision, and population growth), and historical factors (Griffiths and Marjoram, '96). Thus, while coalescent modeling is the most rigorous way to test whether the observed pattern of haplotype diversity departs from neutral expectations, the analysis is complex and beyond the scope of this report. Nevertheless, coalescent simulations carried out in collaboration with P. Andolfatto (manuscript in preparation) do suggest a trend towards a non-neutral excess of linkage disequilibrium in the data, given laboratory-based estimates of the recombination rate.

The degree of haplotype dimorphism is surprising, and may indicate that selection has a pervasive effect on the distribution of variation throughout the genome. Balancing selection could maintain the existence of two haplotype classes at many loci by allowing the buildup of selectively neutral variation linked to sites experiencing strong selection (Hudson et al., '87). There are actually several reports of this phenomenon in relation to Fast/Slow allozyme polymorphisms (see Moriyama and Powell ('96) for references). Qualitative re-examination of the published sequences throughout several of these genes in 100–300 base pair windows also provides evidence for divergent haplotype classes very analogous to our results (results not shown). However, extrapolation of these patterns throughout the genome would imply that there is one site experiencing strong balancing selection every few kilobases, which seems unlikely. It is also possible that there is sufficient epistatic selection between linked sites to maintain higher than expected levels of linkage disequilibrium. Burger and Gimelfarb ('99) have reported results

of a simulation study where multilocus diploid genotype fitnesses were randomly assigned from a normal distribution and populations of infinite size were allowed to evolve to equilibrium. They showed that common polymorphisms can persist at a considerably higher frequency than is typically assumed to occur under mutation-selection balance models. Current methods of quantitative genetic analysis do not have the resolution to test the distribution of multilocus phenotypic, let alone fitness, effects, so it is hard to judge whether such a mechanism may be prevalent enough to support the divergence of haplotype classes throughout the genome.

Finally, a simpler explanation is that modern *D. melanogaster* may actually have formed by admixture of two populations that had been isolated for a large portion of the existence of the species. The combination of recombination and gene conversion must have been sufficient to cause linkage equilibrium between sites separated by more than a few hundred base pairs, as is typically observed, but there may not have been sufficient time to break up associations over shorter stretches. This type of explanation has also been proposed by several authors to explain the prevalence of haplotype dimorphism in *Arabidopsis thaliana* (Hanfstingl et al., '94; Purugganan and Suddith, '99). In principle, restricted migration between subpopulations, possibly with different growth rates, could also explain the data, although Andolfatto and Przeworski (submitted) argue that estimates based on F_{ST} values obtained from existing populations are not consistent with such a simple demographic explanation. Extensive sampling on a global scale throughout the genome will be required to address this issue.

Whatever the explanation, the existence of haplotype clusters throughout the genome, if it is confirmed on larger sample sizes, challenges some of the assumptions of population and quantitative genetic analysis. Most tests of neutrality assume that populations have evolved to a state of equilibrium, but this may not be the case if the time since admixture has been insufficient to break up linkage disequilibrium over these short stretches of DNA. From a quantitative genetic perspective, the existence of linkage disequilibrium over several hundred nucleotides has a major effect on the design of association studies. This is true irrespective of the reasons for the haplotype structure, as it affects the number and distribution of sites that must be surveyed in screens for associations between SNPs and phenotypes. It also compromises the interpretation of whether significant associations are direct or due to tight linkage. In addition, haplotype clades have been proposed as a natural way to overcome some of the problems in multiple comparison testing, because they take advantage of the history of the sample (Templeton et al., '87). The possibility that quantitative genetic effects are embedded in sequences that have only recently been admixed has considerable implications for understanding the reasons for the maintenance of genetic variation in species that show extensive haplotype dimorphism.

ACKNOWLEDGMENTS

We thank David Mindell for access to the ABI 377 at the University of Michigan Museum of Zoology that allowed us to obtain the gene sequences. Initial analysis of the STS data was performed by Salam Bidwan with computer support from Ravonda Pokrzywa and David Bird at NCSU. Thanks to Peter Andolfatto, Michael Purugganan, Trudy Mackay, and Bruce Kimmel for discussions and encouragement. This work was supported by grants from the David and Lucille Packard Foundation to GG, and Rhone-Poulenc Rorer to R.H.

LITERATURE CITED

Begun DJ, Aquadro CF. 1992. Levels of naturally occurring DNA polymorphism correlate with recombination rates in *D. melanogaster*. Nature 356:519–520.

Burger R, Gimelfarb A. 1999. Genetic variation maintained in multilocus models of additive quantitative traits under stabilizing selection. Genetics 152:807820.

Chee M, Yang R, Hubbell E, Berno A, Huang XC, Stern D, Winkler J, Lockhart DJ, Morris MS, Fodor SP. 1996. Accessing genetic information with high-density DNA arrays. Science 274:610–614.

Chen X, Livak KJ, Kwok PY. 1998. A homogeneous, ligation-mediated DNA diagnostic test. Genome Res 8:549–556.

Ewing B, Hillier L, Wendl MC, Green P. 1998. Base-calling of automated sequencer traces using phred: I. Accuracy assessment. Genome Res 8:175–185.

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

Gibson G, van Helden S. 1997. Is function of the Drosophila homeotic gene Ultrabithorax canalized? Genetics 147:1155–1168.

Gibson G, Wemple M, van Helden S. 1999. Potential variance affecting homeotic Ultrabithorax and Antennapedia phenotypes in *Drosophila melanogaster*. Genetics 151:1081–1091.

Gordon D, Abajian C, Green P. 1998. Consed: a graphical tool for sequence finishing. Genome Res 8:195–202.

Griffiths RC, Marjoram P. 1996. Ancestral inference from samples of DNA sequences with recombination. J Comput Biol 3:479–502.

Gurganus M, Nuzhdin S, Leips J, Mackay TFC. 1999. High-resolution mapping of quantitative trait loci for sternopleural bristle number in *Drosophila melanogaster*. Genetics 152:1585–1604.

Hanfstingl U, Berry A, Kellog E, Costa JT 3rd, Rudiger W, Ausubel F. 1994. Haplotypic divergence coupled with lack of diversity at the *Arabidopsis thaliana* alcohol dehydrogenase locus: roles for both balancing and directional selection? Genetics 138:811–828.

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

Kreitman M, Hudson RR. 1991. Inferring the evolutionary histories of the *Adh* and *Adh-dup* loci in *Drosophila melanogaster* from patterns of polymorphism and divergence. Genetics 127:565–582.

Liu J, Mercer JM, Stam L, Gibson G, Zeng ZB, Laurie CC. 1996. Genetic analysis of a morphological shape difference in the male genitalia of *Drosophila simulans* and *D. mauritiana*. Genetics 142:1129–1145.

Long AD, Mullaney SL, Reid L, Fry JD, Langley CH, Mackay TFC. 1995. High resolution mapping of genetic factors affecting abdominal bristle number in *Drosophila melanogaster*. Genetics 139:1273–1291.

Lynch M, Walsh B. 1997. Genetics and analysis of quantitative traits. Sunderland, MA: Sinauer Associates.

Moriyama EN, Powell JR. 1996. Intraspecific nuclear DNA variation in *Drosophila*. Mol Biol Evol 13:261–277.

Nuzhdin SV, Pasyukova E, Dilda C, Zeng ZB, Mackay TFC. 1997. Sex-specific quantitative trait loci affecting longevity in *Drosophila melanogaster*. Proc Natl Acad Sci USA 94:9734–9739.

Pastinen T, Kurg A, Metspalu A, Peltonen L, Syvanen AC. 1997. Minisequencing: a specific tool for DNA analysis and diagnostics on oligonucleotide arrays. Genome Res 7:606–614.

Purugganan MD, Suddith J. 1999. Molecular population genetics of floral homeotic loci. Departures from the equilibrium-neutral model at the APETALA3 and PISTILLATA genes of *Arabidopsis thaliana*. Genetics 151: 839–848.

Ross P, Hall L, Smirnov I, Haff L. 1998. High level multiplex genotyping by MALDI-TOF mass spectrometry. Nat Biotechnol 16:1347–1351.

Saiki RK, Bugawan TL, Horn GT, Mullis KB, Erlich HA. 1986. Analysis of enzymati–cally amplified beta-globin and HLA-DQ alpha DNA with allele-specific oligonucleotide probes. Nature 324:163–166.

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

Tajima F. 1993. Unbiased estimation of evolutionary distance between nucleotide sequences. Mol Biol Evol 10:677–688.

Templeton AR, Boerwinkle E, Sing CF. 1987. A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping: I. Basic theory and an analysis of alcohol dehydrogenase activity in *Drosophila*. Genetics 117:343–351.

Underhill PA., Jin L, Lin A, Mehdi S, Jenkins T, Vollrath D, Davis RW, Cavalli-Sforza LL, Oefner P. 1997. Detection of numerous Y chromosome biallelic polymorphisms by denaturing high-performance liquid chromatography. Genome Res 7:996–1005.

Wang G, Fan JB, Siao CJ, Berno A, Young P, Sapolsky R, Ghandour G, Perkins N, Winchester E, Spencer J, Kruglyak L, Stein L, Hsie L, Topaloglou T, Hubbell E, Robinson E, Mittmann M, Morris MS, Shen N, Kilburn D, Rioux J, Nusbaum C, Rozen S, Hudson TJ, Lander ES. 1998. Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. Science 280:1077–1082.

Weir BS. 1996. Genetic data analysis II. Sunderland, MA: Sinauer Associates.

Werle E, Schneider C, Renner M, Volker M, Fiehn W. 1994. Convenient single-step, one tube purification of PCR products for direct sequencing. Nucleic Acids Res 22: 4354–4355.

Zapata C, Alvarez G. 1993. On the detection of nonrandom associations between DNA polymorphisms in natural populations of *Drosophila*. Mol Biol Evol 10:823–841.

Zeng ZB. 1993. Theoretical basis for separation of multiple linked gene effects in mapping quantitative trait loci. Proc Natl Acad Sci USA 90:10972–10976.