DOI: 10.1534/genetics.105.048207

Nucleotide Polymorphism and Linkage Disequilibrium in Wild Populations of the Partial Selfer *Caenorhabditis elegans*

Asher D. Cutter[1]

Institute of Evolutionary Biology, University of Edinburgh, Edinburgh EH9 3JT, United Kingdom

Manuscript received July 13, 2005
Accepted for publication September 28, 2005

ABSTRACT

An understanding of the relative contributions of different evolutionary forces on an organism's genome requires an accurate description of the patterns of genetic variation within and between natural populations. To this end, I report a survey of nucleotide polymorphism in six loci from 118 strains of the nematode *Caenorhabditis elegans*. These strains derive from wild populations of several regions within France, Germany, and new localities in Scotland, in addition to stock center isolates. Overall levels of silent-site diversity are low within and between populations of this self-fertile species, averaging 0.2% in European samples and 0.3% worldwide. Population structure is present despite a lack of association of sequences with geography, and migration appears to occur at all geographic scales. Linkage disequilibrium is extensive in the *C. elegans* genome, extending even between chromosomes. Nevertheless, recombination is clearly present in the pattern of polymorphisms, indicating that outcrossing is an infrequent, but important, feature in this species ancestry. The range of outcrossing rates consistent with the data is inferred from linkage disequilibrium, using "scattered" samples representing the collecting phase of the coalescent process in a subdivided population. I propose that genetic variation in this species is shaped largely by population subdivision due to self-fertilization coupled with long- and short-range migration between subpopulations.

UNDERSTANDING the genetic basis of evolution requires an accurate description of the patterns of genetic variation in natural populations. The landscape of genetic diversity is molded by the effects of mutation, selection (positive, negative, and balancing), recombination, stochasticity (*i.e.*, genetic drift), and demography. We can attempt to infer how each of these general processes actually contributes to observed natural patterns of diversity by applying the extensive population genetics theory that has developed around the notion of neutral molecular markers and their nonneutral linked loci. Among the factors that can influence patterns of genetic variation in *Caenorhabditis elegans*, its partially selfing breeding system seems likely to play a prominent role. The effect of self-fertilization on diversity is threefold: reduced effective population size and reduced genomewide effective recombination rates, both due to increased homozygosity, and elevated isolation among individuals and subpopulations induced by inbreeding (Charlesworth 2003). Consequently, a predominantly selfing mode of reproduction may be expected to lead to low polymorphism, extensive linkage disequilibrium, and high population subdivision, although migration and metapopulation processes can lead to other patterns (Nordborg 2000; Ingvarsson 2002). Here, I test these predictions by quantifying levels of nucleotide diversity, linkage disequilibrium, and population structure from loci across two chromosomes of 118 individuals in population samples of wild *C. elegans*.

Since the first natural survey of nucleotide variation (Kreitman 1983), most such studies have focused on obligately outcrossing species, such as humans and species of Drosophila (Zhao *et al.* 2000; Yu *et al.* 2001). However, recent large-scale resequencing efforts in plants (*Arabidopsis thaliana* and *Zea mays*) have augmented previous surveys aimed at describing the processes that affect polymorphism throughout the genome of self-fertilizing species (Mitchell-Olds 2001; Nordborg *et al.* 2005; Schmid *et al.* 2005; Wright *et al.* 2005). The patterns of diversity across sequence space and geographic space frequently deviate from neutral predictions, so that population genetic models that include both selection and demographic history are necessary to account for the observed patterns. Like *A. thaliana*, *C. elegans* is capable of close inbreeding by selfing: *C. elegans* reproduce either by hermaphrodite self-fertilization or by hermaphrodite outcrossing with males. Under laboratory conditions, males and outcrossing are infrequent (Hodgkin and Doniach 1997; Chasnov and Chow 2002; Stewart and Phillips 2002; Cutter *et al.* 2003a). Although this is also expected to be true in nature, the breeding system is difficult to characterize quantitatively (Graustein *et al.* 2002; Cutter and Payseur 2003;

Sequence data from this article have been deposited with the EMBL/GenBank Data Libraries under accession nos. DQ231609–DQ232315.

[1]*Address for correspondence:* Institute of Evolutionary Biology, University of Edinburgh, W. Mains Rd., King's Bldgs., Ashworth Labs, Edinburgh EH9 3JT, United Kingdom. E-mail: asher.cutter@ed.ac.uk

Denver *et al.* 2003; Barrière and Félix 2005; Haber *et al.* 2005), and further, it is only beginning to become clear what role migration plays in structuring genetic variation in this species (Barrière and Félix 2005).

Despite the extensive literature on the nematode *C. elegans* as a model for many aspects of biology, *C. elegans* as a model for population genetics is in its nascent stages (Delattre and Félix 2001). A number of studies have investigated genetic variation in this species, using several types of molecular markers (Thomas and Wilson 1991; Koch *et al.* 2000; Graustein *et al.* 2002; Denver *et al.* 2003; Jovelin *et al.* 2003; Sivasundar and Hey 2003; Barrière and Félix 2005; Haber *et al.* 2005; Sivasundar and Hey 2005; Stewart *et al.* 2005). However, all but three very recent studies have had to rely on a haphazard assortment of strains. Without suitable sampling, characterization of intra- and interpopulation statistics is not possible. Estimates of nucleotide diversity (*e.g.*, π and θ) have been published for only four nuclear loci (Graustein *et al.* 2002; Jovelin *et al.* 2003), with no explicit sampling within local populations, and formal analyses of linkage disequilibrium and population structure are limited (Koch *et al.* 2000; Sivasundar and Hey 2003; Barrière and Félix 2005; Haber *et al.* 2005). The contrast is striking with *C. elegans*' position at the forefront of approaches used to estimate other parameters relevant to molecular evolution and population genetics, such as mutation rate (Denver *et al.* 2004).

Here I report DNA sequence variation from six loci on two chromosomes (II and X) in 106 strains of *C. elegans* from three European countries plus 12 worldwide strains from the Caenorhabditis Genetics Center. I quantify diversity in these population samples and describe the linkage disequilibrium and population structure. The results show low silent nucleotide diversity, both within and between populations, and linkage disequilibrium within and between chromosomes in the European samples. Nevertheless, within-population diversity is only moderately reduced relative to that of the species as a whole, and there is clear evidence for recombination and migration between populations.

MATERIALS AND METHODS

Nematode populations: For this study, I isolated DNA from 118 isohermaphrodite strains (supplemental Table 1 at http://www.genetics.org/supplemental/), including 23 German strains (Haber *et al.* 2005), 57 strains from France (Barrière and Félix 2005), and 12 strains from the Caenorhabditis Genetics Center (CGC) with worldwide distribution: N2, AB1, AB4, CB4853, CB4854, CB4856, CB4857, CB4858, KR314, RC301, RW7000, and TR403. These CGC strains were originally isolated in England, Australia, the United States, Canada, Germany, and France (Hodgkin and Doniach 1997). In addition, 26 new strains were isolated from single wild individuals around Edinburgh, Scotland (supplemental Table 1 at http://www.genetics.org/supplemental/). The sampling sites in Scotland include compost bins in two public allotment gardens (Midmar 1-39 and 2-43, West Mains) and discarded compost from a mushroom farm in North Berwick. One isolate from North Berwick derives from a postproduction mushroom growth flat infested with nematodes to an extent that the entire surface ($\sim$0.1 m^2) could be seen glistening with crawling worms (estimated $\geqslant$10,000 individuals); many such flats were present in the darkhouse. I refer to the strains obtained from the CGC as "CGC strains" and strains derived from wild population samples as "European strains." The protocol for the Scottish nematode isolations was kindly provided by M. Félix (personal communication). Briefly, small samples of compost ($\sim$2 ml), or individual isopods (crushed, nine *C. elegans* strains from *Porcellio scaber*, *P. spinicornis*, and an unidentified species), were placed on standard 6-cm NGM-lite agar plates spotted with *Escherichia coli* OP50. After $\sim$4 hr, individual nematodes were isolated on separate NGM-lite agar plates. Self-fertile individuals were inspected under 100$\times$ microscopy for morphological characters. Progeny of candidate Caenorhabditis were subjected to mating tests and species identity was confirmed when mating trials with N2 or CB4856 males generated $\sim$50% male offspring. Strains derived from single wild individuals were subsequently propagated by intermittent transfer to new agar plates with strain name designations ED3000, ED3005, ED3006, ED3008, and ED3010–ED3031.

Molecular methods: DNA from pooled samples of five worms from each iso-hermaphrodite strain was isolated using a NaOH digestion protocol (Floyd *et al.* 2002). Heterozygote detection is a formal possibility with this approach, although the small DNA pools coupled with multiple generations of inbreeding in the laboratory make it unlikely; no heterozygotes were detected. I selected three loci for sequencing on each of *C. elegans* chromosomes II and X (Figure 1), choosing genes that contained an intron >500 bp long and that are distributed across most of the map length of each chromosome. Forward and reverse primers for both amplification and sequencing were designed from the Wormbase genome sequence in coding regions spanning a long intron (Table 1). Both strands were sequenced on an ABI Prism 3730 automatic sequencer. Sequence data from this article have been deposited in GenBank under accession nos. DQ231609–DQ232315. Feature statistics for each locus were obtained from Wormbase release 143 (www.wormbase.org).

Sequence alignment and analysis: Sequences were aligned in Sequencher v. 4.0 followed by manual adjustment in BioEdit and removal of primer sequence. Sequence data analyses (diversity from pairwise differences π, diversity from the number of segregating sites θ, tests of neutrality, tests of population structure, linkage disequilibrium, and recombination) were performed using DnaSP v. 4.1 (Rozas *et al.* 2003), RecMin (Myers and Griffiths 2003), and LIAN v. 3.1 (Haubold and Hudson 2000). Sites corresponding to indels or incomplete data were excluded from the analyses. Consequently, the French strain JU406 was excluded in analyses of concatenated sequence data and analyses of locus E01G4.6 because no amplification product or sequence was obtained for this locus. This may lead to slight underestimation of diversity if a mutation in the primer region is responsible for the amplification failure. Because the presence of indels relative to N2 makes it problematic to assign absolute genomic positions to variable sites, the positions indicated in the table of polymorphism in Figure 1 correspond to unique locations within a concatenated sequence alignment. Coalescent simulations were implemented in DnaSP to test for significant differences in diversity levels, and the program *Q*-value was used to compute false-discovery rates (Storey and Tibshirani 2003). Neighbor-joining trees were constructed with concatenated

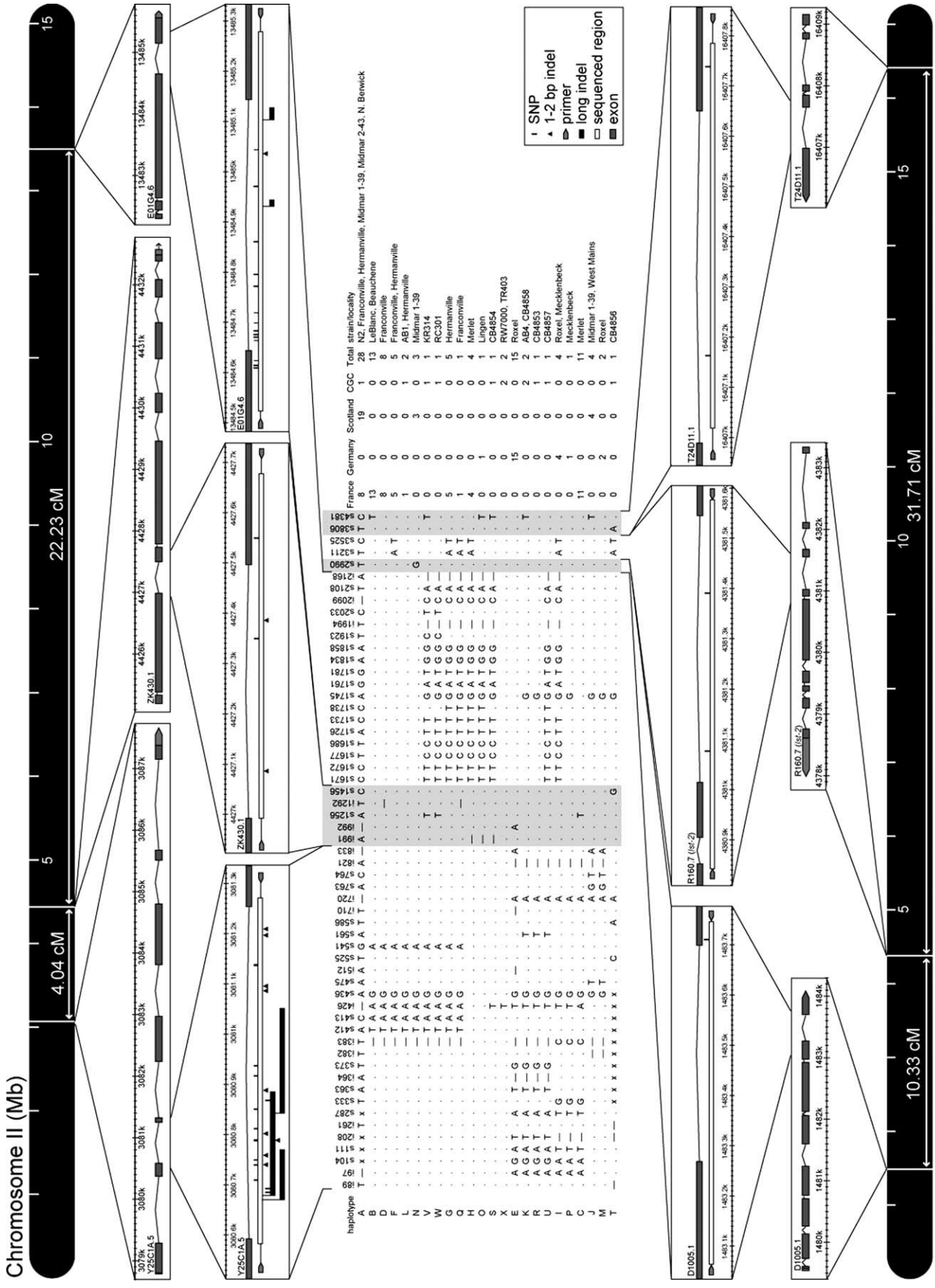

FIGURE 1.—Summary of sequenced regions, polymorphic sites, and haplotypes.

TABLE 1

Loci sequenced in *C. elegans*

Locus	Chromosome	Position (cM)	Intron	Forward primer	Reverse primer	Product (bp)	Intron (bp)	Exon (bp)
Y25C1A.5	II	−8.58	2	TCAAGTCGGATCTTTTGGAGA	GGCTGGTGATGATGACTGAA	797	661	136
ZK430.1	II	−4.54	2	GTTGTCGGTGTTGGAGGAGT	TGCAAAGAGAGCAGCAAGAA	798	504	294
E01G4.6	II	17.69	3	CAGCAATGCGACAGGAAGTA	CACGTCAGATGGTTGGACTG	836	501	335
D1005.1	X	−18.05	7	GTCGGGAGCCCATAACACTA	TGGGCAAACTCCTTCTTGTC	724	430	294
R160.7	X	−7.77	3, 4	GATTCTGCACGACACATTGC	ACTGGGTTGTAGGCAGATGG	776	581	195
T24D11.1	X	23.94	4	CGTGTGGGGAGAGCTTATTC	GGGTAGGTACGGGGAATGTC	887	662	225

sequences using PAUP* v. 4.0 and manipulated in TreeView v. 1.6.6. Because of the evidence for recombination in these data (see below), the trees resulting from concatenated sequence haplotypes should not be used to infer phylogenetic relationships between strains. The program Structure 2.0 (Pritchard *et al.* 2000) was used to infer a maximum-likelihood estimate for the number of subpopulations represented in the European and CGC strains, on the basis of the average of triplicate runs for values of K subpopulations from 1 to 25 (using the admixture model and independent allele frequencies).

To better approximate the assumptions of the neutral coalescent by analyzing samples in the "collecting phase" of a population with structure (Nordborg 1997; Wakeley 1999; Wakeley and Lessard 2003; Lessard and Wakeley 2004), in some analyses I employed a resampling scheme to create 1000 random subsets of strains composed of a single individual from each European and CGC sampling locality. The ability of this "scattered sample" approach to truly approximate the neutral coalescent process depends on how well *C. elegans* conforms to the assumptions of an island model of migration connecting many demes (Wakeley 1999; Wakeley and Lessard 2003; Lessard and Wakeley 2004). The available data suggest that these assumptions may be approximately correct, although our understanding of the scale of population subdivision is decidedly imperfect. From the "scattered" random samples, linkage disequilibrium was measured for pairs of sites between loci and between chromosomes, using the squared correlation between pairs of sites (r^2) in the RSQ application of libsequence (Thornton 2003). The resulting mean r^2-values (and 2.5 and 97.5 percentiles) were then evaluated according to the equation given in the discussion to infer outcrossing rates given a point estimate of N_e (see results) and the recombination distances (c) between locus pairs (Figure 1; Table 1). This scattered-sample approach was also used to estimate π_{si} and θ_{si} with polydNdS (Thornton 2003) and the population recombination parameter for each chromosome (ρ_e) with the LDhat program pairwise (Fearnhead and Donnelly 2001).

RESULTS

DNA polymorphism: Figures 1 and 2 and Table 2 summarize nucleotide polymorphism in the six gene regions surveyed on chromosomes II and X. In a total of 3372.4 bp of silent sites across all regions, the levels of per-site variation for the European strains are $\pi_{si} = 0.00215$ and $\theta_{si} = 0.00159$. Of the 28 total segregating sites in these strains, 23 are located in introns and 5 at synonymous coding sites. Diversity at synonymous sites ($\pi_{syn} = 0.00629$ and $\theta_{syn} = 0.00575$) is only nominally higher than that for all silent sites ($P > 0.05$), although very few synonymous sites were included in this study (166.4 bp). No polymorphisms were detected in the 562.6 bp of nonsynonymous sites. The variants include 16 transitions and 12 transversions, yielding a transition: transversion ratio (ts/tv) of 1.33, somewhat lower than previous reports based on mutation-accumulation lines and comparisons between CGC strains (Koch *et al.* 2000; Denver *et al.* 2003, 2004). The different loci are highly heterogeneous in their diversity levels, with π_{si} varying >50-fold among regions and θ_{si} varying 17-fold. Such variation among loci is not unexpected, given the potential influences of different local mutation rates, selection, demography, and stochasticity associated with low polymorphism. Nucleotide diversity appears lower on the X chromosome, but this is not a significant difference (Wilcoxon $P > 0.1$). In addition to single-nucleotide polymorphisms, the sequenced region of locus Y25C1A.5 contained a high-frequency 206-bp indel variant (containing an additional 3 polymorphic sites and one variable length repeat not included in analyses) and locus E01G4.6 contained two indels 12 and 24 bp in length (Figure 1). Also in locus Y25C1A.5, the Hawaiian strain CB4856 contains two long deletions (100 and 209 bp) that largely overlap the indel region observed in the other strains. Short indels of one or two nucleotides, generally associated with simple repeats, were present in three loci (eight in Y25C1A.5, two in ZK430.1, and one in E01G4.6). None of the loci on the X chromosome contained indels or variable-length repetitive sequences.

To compare the levels of polymorphism found in other studies with these wild population samples, I also evaluated genetic differences among 12 strains obtained from the Caenorhabditis Genetics Center (CGC strains) that have been included in previous studies (Hodgkin and Doniach 1997; Wicks *et al.* 2001; Graustein *et al.* 2002; Denver *et al.* 2003; Jovelin *et al.* 2003; Sivasundar and Hey 2003; Haber *et al.* 2005). This geographically broader sample of strains shows significantly higher diversity than the European samples for some loci ($P \leq 0.02$ for π_{si} and θ_{si} of ZK430.1, E01G4.6, and T24D11.1), but is only nominally higher for all sequences considered together ($P = 0.14$; Tables 3 and 4). In these CGC strains, the 29 single-nucleotide polymorphisms (SNPs)

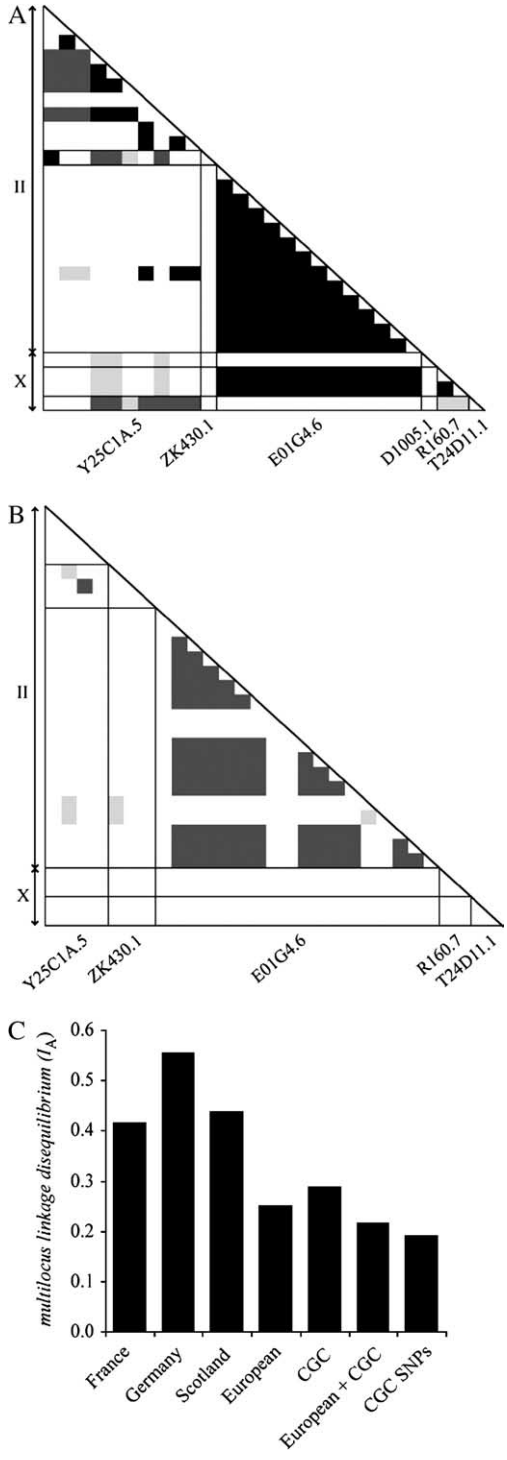

FIGURE 2.—Linkage disequilibrium in *C. elegans* European strains (A) and CGC strains (B). Different loci are demarcated by lines and chromosomes are indicated on the left. Shadings indicate Fisher's exact test significance levels: light shading, $P < 0.05$; medium shading, $P < 0.01$; dark shading, significant after Bonferroni correction. (C) Multilocus linkage disequilibrium levels for the standardized index of association (I_A) for different collections of strains (all $P < 0.001$).

exhibited a ts/tv ratio of 1.64, similar to previous reports (KOCH *et al.* 2000; DENVER *et al.* 2003, 2004).

Estimates of diversity yield an average European effective population size (N_e) estimate for *C. elegans* of $\sim 5 \times 10^4$ (for CGC strains, $N_e \sim 9 \times 10^4$), given a per-site neutral mutation rate of 9.0×10^{-9} and assuming that equilibrium has been reached (*i.e.*, $\theta = 4 N_e \mu$) (DENVER *et al.* 2004; KEIGHTLEY and CHARLESWORTH 2005). When estimated per locus, N_e varies between ~7000 and ~160,000 for European strains and up to ~360,000 for the CGC strains. However, it may be most appropriate to calculate N_e from a set of strains that includes only a single individual from each subpopulation to better approximate the assumptions of the neutral coalescent process (LESSARD and WAKELEY 2004). Estimating N_e from mean π_{si} (0.00295) or θ_{si} (0.00276) derived using the scattered-sampling approach yields $N_e \sim 8 \times 10^4$. These values of global population size are substantially higher than N_e inferred from microsatellites and AFLP data for local populations (SIVASUNDAR and HEY 2003; BARRIÈRE and FÉLIX 2005), but are still rather small relative to the high census densities that nematodes can achieve.

Linkage disequilibrium and recombination: For the European samples, intralocus linkage disequilibrium is strong within the three loci that contain more than one segregating site (Figure 2, supplemental Figure 1 at http://www.genetics.org/supplemental/). In addition, interlocus linkage disequilibrium occurs both within and between chromosomes. After correction for multiple tests, 30% of all pairs of sites show significant linkage disequilibrium. Due to these high levels of linkage disequilibrium coupled with the low polymorphism, only 14 haplotypes (h) are present in the entire sample of 106 strains from France, Germany, and Scotland or 16 including indels in the construction of haplotypes (Figure 1; Table 4). An additional 8 haplotypes are found among the CGC strains. The most common haplotype, present in a minority of French and a majority of Scottish samples, is identical to that of the canonical strain N2, which was originally isolated in Bristol, England (Figure 1). The extensive linkage disequilibrium is not a consequence of regional population structure alone, since very similar patterns are observed for pooled European samples and for samples from each country analyzed separately (Figure 2; *cf.* supplemental Figure 1 at http://www.genetics.org/supplemental/).

Much weaker linkage disequilibrium is seen in the 12 CGC strains than for the European samples (Figure 2). Measures of overall linkage disequilibrium differ significantly from the neutral expectation ($P < 0.001$), but no pairs of sites are significantly associated after a multiple-tests correction (Figure 2). These results are generally consistent with an analysis of linkage disequilibrium based on microsatellites in CGC strains (SIVASUNDAR and HEY 2003). I also calculated linkage disequilibrium levels among 230 SNPs scored in a different set of 11

TABLE 2

Summary of haplotypes, diversity, and population differentiation for European population samples

Locus (chromosome)	Silent sites	S	h	H_d	Indels	Diversity: π_{si}	Diversity: θ_{si}	D_{ST}[a]	D_{ST}[b]	F_{st}[a]	Nm[a]	F_{st}[b]	Nm[b]
Y25C1A.5 (II)	657.8	10	5	0.749	12	0.00435	0.00290	0.00137	0.00217	0.418	0.35	0.531	0.22
ZK430.1 (II)	534.5	1	2	0.188	2	0.00035	0.00036	−0.00001	0.00020	0.182	1.13	0.556	0.20
E01G4.6 (II)	506.4	13	4	0.362	3	0.00641	0.00491	0.00038	0.00083	0.078	2.95	0.160	1.31
D1005.1 (X)	462.8	1	2	0.056	0	0.00012	0.00041	−0.00002	0.00001	0.080	2.88	0.154	1.38
R160.7 (X)	550.8	2	2	0.297	0	0.00108	0.00069	0.00027	0.00028	0.115	1.93	0.272	0.67
T24D11.1 (X)	660.0	1	2	0.285	0	0.00043	0.00029	0.00007	0.00028	0.044	5.42	0.880	0.03
Concatenated	3372.4	28	14	0.877	17	0.00215	0.00159	0.00039	0.00065	0.235	0.82	0.371	0.42
Average	562.1	4.67	2.83	0.323	2.8	0.00212	0.00159	0.00035	0.00063	0.153	2.44	0.425	0.64

S, segregating sites; h, haplotypes; H_d, haplotype diversity.
[a] Assuming three populations on the basis of country of origin (France, Germany, Scotland).
[b] Assuming seven populations on the basis of localities with more than two individuals.

CGC strains by Koch *et al.* (2000). These SNP data indicate extensive linkage disequilibrium within and between chromosomes (Figure 3), although, because of the large number of comparisons, none is individually significant after correction with Bonferroni or false-discovery rate procedures (Storey and Tibshirani 2003). A measure of multilocus linkage disequilibrium (standardized I_A) (Agapow and Burt 2001) for the SNP data set is comparable to what was found for the CGC strains (Figure 2C). Again using the scattered-sample approach to better approximate neutral processes in a subdivided population (taking a single individual from each European and CGC locality), average pairwise linkage disequilibrium (r^2) between loci varies from 0.007 to 0.23 and r^2 between chromosomes averages 0.08.

Despite the extensive linkage disequilibrium among sites, recombination is detectable. A conservative measure of the minimum number of recombination events for these data, based on the four-gamete test (Hudson and Kaplan 1985), yields a value of $R_m = 2$ for the 106 European strains. Myers and Griffiths' (2003) method generates a lower bound of $R_h = 5$ for the number of recombination events in this data set, with recombination between loci predicted to have occurred within and between both chromosomes II and X. Including the CGC strains raises R_m to 3 and R_h to 9, with both intrachromosomal and interchromosomal recombination events predicted to have occurred among the 12 CGC strains alone ($R_m = 2$, $R_h = 4$). The markers used here cover 23% of the map length of the genome, so scaling up the estimated minimum number of recombination events suggests values of at least 22.0 European and 39.5 worldwide recombination events in the history of the genome of these strains since their most recent common ancestor. Recombinant haplotypes are also evident in the 230 SNPs scored in 11 CGC strains by

TABLE 3

Diversity levels per locus by country of origin

	Diversity: π_{si}				Diversity: θ_{si}				
Locus	France	Germany	Scotland	CGC	France	Germany	Scotland	CGC	Reference
Y25C1A.5[a]	0.00326	0.00287	0.00281	0.00505	0.00163	0.00288	0.00277	0.00361	This study
ZK430.1	0.00059	0	0.00050	0.00175	0.00040	0	0.00048	0.00182	This study
E01G4.6[b]	0.00767	0.00993	0.00049	0.01480	0.00559	0.00778	0.00047	0.01142	This study
D1005.1	0	0	0.00043	0	0	0	0.00054	0	This study
R160.7	0.00143	0.00101	0	0.00054	0.00079	0.00091	0	0.00108	This study
T24D11.1	0.00054	0.00013	0.00041	0.00099	0.00033	0.00041	0.00040	0.00100	This study
glp-1[c]	—	—	—	0.00140	—	—	—	—	Graustein *et al.* (2002)
odr-3[d]	—	—	—	0.00011	—	—	—	—	Jovelin *et al.* (2003)
spe-9[e]	—	—	—	0.00050	—	—	—	—	Graustein *et al.* (2002)
tra-2[c]	—	—	—	0	—	—	—	—	Graustein *et al.* (2002)

[a] CB4856 excluded.
[b] JU406 excluded.
[c] $n = 20$.
[d] $n = 10$.
[e] $n = 16$.

TABLE 4

Diversity in each sample locality for concatenated sequence

Country	Location	n	S	h	H_d	Diversity: π_{si}	Diversity: θ_{si}
France	Franconville	12	19	3	0.530	0.00121	0.00184
France	Hermanville[a]	12	21	4	0.697	0.00326	0.00203
France	LeBlanc	12	0	1	0	0	0
France	Merlet	19	23	3	0.608	0.00256	0.00193
Germany	Roxel	19	19	3	0.374	0.00139	0.00181
Scotland	Edinburgh (1-39)	14	11	3	0.560	0.00100	0.00098
Scotland	Edinburgh (2-43)	8	0	1	0	0	0
Scotland	Edinburgh (Midmar)	22	11	3	0.394	0.00068	0.00086
France	All[a]	56	22	7	0.852	0.00222	0.00142
Germany	All	23	25	5	0.557	0.00224	0.00195
Scotland	All	26	11	3	0.446	0.00082	0.00082
European	All[a]	105	28	14	0.877	0.00215	0.00159
Worldwide	CGC	12	29	9	0.939	0.00329	0.00290

S, segregating sites; *h,* haplotypes; H_d, haplotype diversity.
[a] JU406 excluded.

Koch *et al.* (2000): at least $R_m = 24$ ($R_h = 26$) recombination events are estimated in the history of the sample.

Population structure: For analyses of European population structure, samples were partitioned either by country of origin or by locality (for localities with more than two animals sampled). Most haplotypes are endemic to a single country, but most polymorphic sites are found in multiple populations (Figure 1). Measures of population structure also provide evidence for some differentiation, with values of F_{st} averaging 0.15 for different loci among countries and 0.43 among localities (Table 2). However, low within-population diversity can lead to nonzero F_{st}-values even in the absence of population structure, making it useful to consider diversity statistics that are not inflated by low polymorphism, such as D_{ST}, the difference between total (π_T) and mean within-population diversity (π_S) (Charlesworth *et al.* 1997; Pannell and Charlesworth 1999). For most of these loci, D_{ST} is very low (Table 2), consistent with the F_{st}-results showing that a higher proportion of all polymorphism is present within populations. Analyses with the program Structure 2.0 (Pritchard *et al.* 2000) suggest that $K = 16$ subpopulations are present in the collection of all strains from Europe and the CGC, although the presence of selfing may make Structure an unreliable method for determining the maximum-likelihood number of subpopulations (D. Falush, personal communication).

In addition, the clustering of haplotypes by genetic distance does not correlate with the country of origin and no fixed differences are present between samples from different countries, indicating that geographic structure is limited (Figure 4). No evidence of isolation by distance is present, on the basis of the lack of a correlation between pairwise F_{st} and rank-order distances between localities (Wilcoxon $P > 0.2$). To the extent that *C. elegans* population dynamics make it appropriate

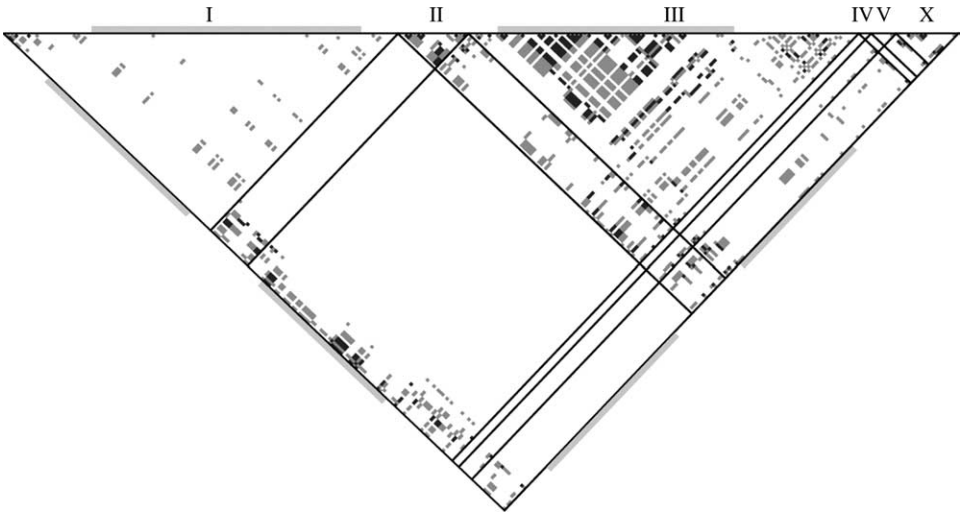

Figure 3.—Linkage disequilibrium between SNPs of *C. elegans.* Sites from different chromosomes are demarcated by lines. Shadings indicate Fisher's exact test significance levels: light shading, $P < 0.05$; dark shading, $P < 0.01$. Rectangles with light shading along the perimeter indicate short stretches of sequence on chromosomes I and III with many SNPs (Koch *et al.* 2000).

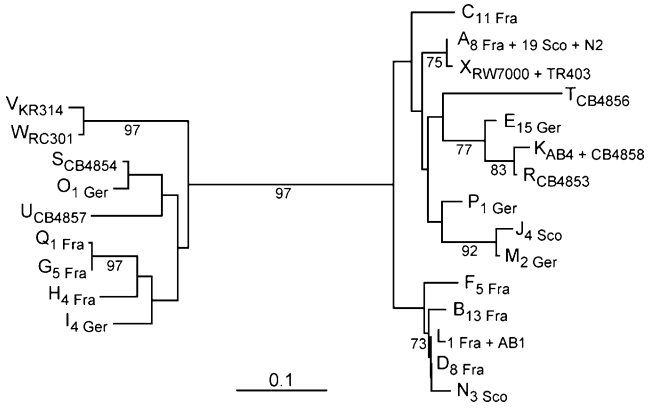

FIGURE 4.—Unrooted neighbor-joining tree for haplotypes derived from a concatenated sequence of 105 European and 12 CGC strains of *C. elegans*. Subscripts indicate the number of strains per haplotype and strain origin (Fra, France; Ger, Germany; Sco, Scotland). Bootstrap values >70% are indicated below the branches. Haplotype designations are as in Figure 1.

to estimate the migration parameter *Nm*, values of *Nm* are not negligible, averaging $Nm = 0.64$ (Table 2). Despite the genetic differentiation, local subpopulations (at the level of both country and locality) harbor levels of genetic variation nearly as high as that observed for all strains (Tables 3 and 4). Only strains from one Scottish locality and the sample from LeBlanc have no variants, and overall the Scottish populations have lower diversity than other European strains ($P = 0.032$; Table 3).

Tests of neutrality: At equilibrium under a standard neutral model, we expect approximately equal values for measures of genetic variation that are based on the number of segregating sites (θ) or on pairwise differences (π) (WATTERSON 1975; NEI and LI 1979; TAJIMA 1983). Statistics such as Tajima's *D* quantify departures from this neutral expectation, and values different from zero suggest the action of nonneutral demographic or selective processes (TAJIMA 1989). The nominal values of TAJIMA's (1989) *D* and FU and LI's (1993) *D** are positive for most loci when all European samples are considered together (but none are significant; Table 5), suggesting an excess of intermediate-frequency polymorphisms. However, when there is population structure, intrapopulation estimates of *D* and *D** are preferable because subdivision in a sample inflates *D*-values (PANNELL 2003). For the French, German, and Scottish samples separately, nearly all values of *D* and *D** are again positive (Table 5), with significant departures from the neutral expectation for two loci among the French samples (for Y25C1A.5 $D = 2.35$, $P < 0.05$; for E01G4.6 $D^* = 1.51$, $P < 0.05$). The German samples for locus T24D11.1 provide the only case with marked negative values of *D* and *D**. At the scale of individual sampling of localities within a country, however, the balance shifts toward slightly negative values of *D* and *D** for most comparisons; in fact, samples from Franconville depart significantly from neutral expectation for E01G4.6 ($D = -2.10$, $P < 0.05$; $D^* = -2.62$, $P < 0.05$). Nevertheless, even at this very local scale, *D* and *D** are significantly positive at Y25C1A.5 and E01G4.6 in some localities.

DISCUSSION

Nucleotide diversity in *C. elegans*: Natural population samples of *C. elegans* from Europe are characterized by low levels of silent-site nucleotide diversity, averaging $\pi_{si} = 0.2\%$. While different loci show substantial variation around this mean, the nucleotide diversity estimates in

TABLE 5

Neutrality tests (Tajima's *D*) per locus within each locality

Country	Location	*n*	Y25C1A.5	ZK430.1	E01G4.6	D1005.1	R160.7	T24D11.1	Average
France	Franconville	12	0.83		−2.10*		−1.45		−0.91
France	Hermanville[a]	13 (12)	2.28*		2.58*		1.88		2.25
France	LeBlanc	12							
France	Merlet	19	1.06	1.47	0.85		1.91		1.32
Germany	Roxel	19	−0.41		−0.99		−0.73		−0.71
Scotland	Edinburgh (M 1-39)	14	0.29	−0.34	−0.34	0.32		−0.34	−0.08
Scotland	Edinburgh (M 2-43)	8							
Scotland	Edinburgh	22	−0.49	−0.64	−0.64	−0.17		−0.64	−0.52
France	All[a]	57 (56)	2.35*	0.60	1.09		1.42	0.85	1.26
Germany	All	23	−0.01		0.98		0.24	−1.16	0.01
Scotland	All	26	0.04	0.05	0.05	−0.31		0.05	−0.02
European	All[a]	106 (105)	1.27	−0.02	0.81	−0.80	0.84	0.56	0.44
CGC	All[b]	12 (11)	1.62	−0.13	1.29		−1.45	−0.05	0.26

* $P < 0.05$.
[a] JU406 excluded from E01G4.6.
[b] CB4856 excluded from Y25C1A.5.

different subpopulations and to the species as a whole are remarkably similar; *i.e.*, most diversity occurs within rather than between populations. The average silent-site diversity estimate for a worldwide sample of CGC strains is somewhat higher than that previously reported ($\pi_{si} =$ 0.33% here *vs.* 0.075% in the literature; Table 3), although the ranges overlap and each study analyzed different strains (Graustein *et al.* 2002; Jovelin *et al.* 2003). Diversity estimated from the CGC strains tends to be higher than estimates from the European population samples, for the same loci, probably reflecting a greater number of sampling localities. Multiple lines of evidence, from different classes of molecular marker, now point to a pattern of both low global and local diversity in *C. elegans* (Sivasundar and Hey 2003; Barrière and Félix 2005; Haber *et al.* 2005). For comparison, autosomal synonymous-site diversity of 1.6% in *Drosophila melanogaster* is ~5 times greater than that in *C. elegans* (Andolfatto 2001) and diversity in the dioecious *C. remanei* is ~10 times greater than that for *C. elegans* (Graustein *et al.* 2002). Global human genetic diversity, on the other hand, is only ~0.08% relative to 0.33% in *C. elegans* (Zhang 2000; Yu *et al.* 2001).

What other genomic factors might contribute to variation in levels of diversity? Introns on the X chromosome show particularly low variation, although it is not clear whether this reflects a real difference, given only three loci per chromosome. Provided that the selfing rate in this species is high, most individuals will be hermaphrodites (XX), so autosomes and the X chromosome will have equivalent effective sizes. Thus, it is unnecessary to adjust diversity levels as in dioecious species to compensate for a different X effective population size. In comparisons with *C. briggsae*, the X generally shows much greater synteny and fewer rearrangements than the autosomes (Stein *et al.* 2003) and nonsynonymous sites (but not synonymous sites) on the X chromosome diverge more slowly than autosomal ones (Cutter and Ward 2005). The nucleotide polymorphism data show an apparent consistency with these observations by having a trend of lower diversity on the X, but an explanation is not clear cut. The loci surveyed for polymorphism here also vary in their local recombinational environment, which in *C. elegans* correlates with SNP density (Cutter and Payseur 2003). On the basis of the recombination rate estimates of Cutter and Payseur (2003), diversity increases with recombination rate (Spearman's $\rho = 0.77$ for θ_{si}, $P = 0.072$; Table 6). With these data alone, one cannot determine whether such a pattern is due to mutational processes that correlate with recombination or to selection at linked sites reducing diversity in low recombination regions (Charlesworth *et al.* 1993; Marais *et al.* 2001, 2004). Other potential factors, such as base composition and *C. elegans*–*C. briggsae* synonymous-site divergence (as a proxy for mutation rate), show no association with the observed levels of diversity ($P > 0.1$). It remains to be tested whether demographic or selective scenarios might also explain variation in levels of polymorphism among loci.

TABLE 6

Features of the sequenced loci

Gene (chromosome)	K_A[a]	δ_S[a,b]	F_{op}[c]	Fraction G + C[d]	R[e] (cM/Mb)
Y25C1A.5 (II)	0.071	1.59	0.456	0.273	2.65
ZK430.1 (II)	0.105	1.63	0.439	0.345	2.32
E01G4.6 (II)	0.107	1.62	0.578	0.463	6.81
D1005.1 (X)	0.055	1.77	0.526	0.431	3.19
R160.7 (X)	0.109	1.19	0.359	0.360	2.91
T24D11.1 (X)	0.033	1.25	0.415	0.377	1.93
Average	0.080	1.51	0.462	0.375	3.30

[a] Nonsynonymous (K_A) and synonymous (δ_S) site substitution rates from Cutter and Ward (2005) comparisons with *C. briggsae*.
[b] Adjusted for correlation with codon usage bias.
[c] Fraction of optimal codons.
[d] Base composition based only on sequenced region.
[e] Recombination rates from Cutter and Payseur (2003).

Linkage disequilibrium and population structure: Linkage disequilibrium within and between loci pervades the *C. elegans* genome. Extensive linkage disequilibrium is found within populations, even between chromosomes, indicating similar ancestries between freely recombining portions of the genome. These results are consistent with the patterns observed for microsatellites and AFLPs within German and French *C. elegans* populations (Barrière and Félix 2005; Haber *et al.* 2005) and with the SNP study of Koch *et al.* (2000), for which I present a formal analysis of linkage disequilibrium. Correspondingly, *C. elegans* genetic diversity is distributed into relatively few haplotypes. The topology of a neighbor-joining haplotype tree reveals two principal groups of haplotypes separated by a long branch, as was also observed for mitochondrial and nuclear sequences in CGC strains (Denver *et al.* 2003). Because of the lack of an appropriate outgroup (silent sites are saturated with differences relative to the congeners *C. briggsae* and *C. remanei*), one cannot reliably infer ancestral states of the polymorphic sites and haplotypes. A pattern of two relatively closely related groups separated by long branches is expected for neutral coalescent trees under selfing (Charlesworth 2003; Hein *et al.* 2004); however, whether the root lies along the long branch in the observed topology is a matter of speculation. It is also important to recognize that the relationship between strains is not strictly tree-like, because recombination, even if rare, causes different portions of the genome of a given strain to have different genealogies (Nordborg 2000).

Most European sampling locations harbor similar levels of polymorphism, with the diversity composed of different combinations of the same variants in each locality or country. Interestingly, the relationships between

the country-specific haplotypes show no strong signature of geographic structure. A lack of geographic structure to *C. elegans* genetic data also has been noted in previous studies (Denver *et al.* 2003; Sivasundar and Hey 2003; Barrière and Félix 2005). The weak geographic structure of the *C. elegans* genetic data coupled with F_{st}-derived values of the migration parameter $Nm > 1$ indicate that migration is a regular occurrence in this species. These observations are consistent with coalescent theory, which predicts that geographic structure should be absent in a large metapopulation (Wakeley and Aliacar 2001).

Despite the strong linkage disequilibrium and haplotype structure in the samples, the pattern of polymorphisms also shows evidence for recombination within and between chromosomes. This result provides strong support for occasional outcrossing in *C. elegans*. How does this evidence of recombination translate into outcrossing rate? One can estimate the outcrossing rate $(1 - s)$, where s is the selfing rate, from linkage disequilibrium in the following way. The outcrossing rate is related to the effective recombination rate by $c_e = c(1 - F)$, where c is the recombination rate and the inbreeding coefficient $F = s/(2 - s)$ (Pollak 1987; Dye and Williams 1997; Nordborg 1997, 2000). In turn, linkage disequilibrium can be predicted in terms of the recombination rate. Assuming that the population has reached equilibrium, the squared correlation coefficient between pairs of sites (r^2) as an estimator of linkage disequilibrium is described by $r^2 \cong 1/(1 + 4 N_e c_e)$, with c_e a function of F and s as above (Hill and Robertson 1968). Solving for $(1 - s)$ yields an estimator of the outcrossing rate:

$$(1 - s) \cong \frac{1 - r^2}{r^2(1 + 8N_e c) - 1}.$$

The genealogy of a structured or partially selfing population can be described as having two phases, in which the "collecting" phase of interdemic relationships is expected to conform to the standard neutral coalescent process (Nordborg 1997; Wakeley 1999; Wakeley and Lessard 2003; Lessard and Wakeley 2004). The effects of population structure are expected to be removed for scattered samples of neutral sites taken from the collecting phase, resulting in a sample subject to other processes that may affect neutral sites in a panmictic population (Wakeley 1999; Wakeley and Lessard 2003; Lessard and Wakeley 2004). Consequently, to best represent the collecting phase of the genealogy, I calculated the mean r^2 for all pairs of sites for each interlocus comparison and for all pairs of sites from different chromosomes (*i.e.*, $c = 0.5$) for random subsets of strains that include a single individual from each European and CGC sampling locality. The appropriateness of this scattered sample approach depends on how well *C. elegans* conforms to the assumptions

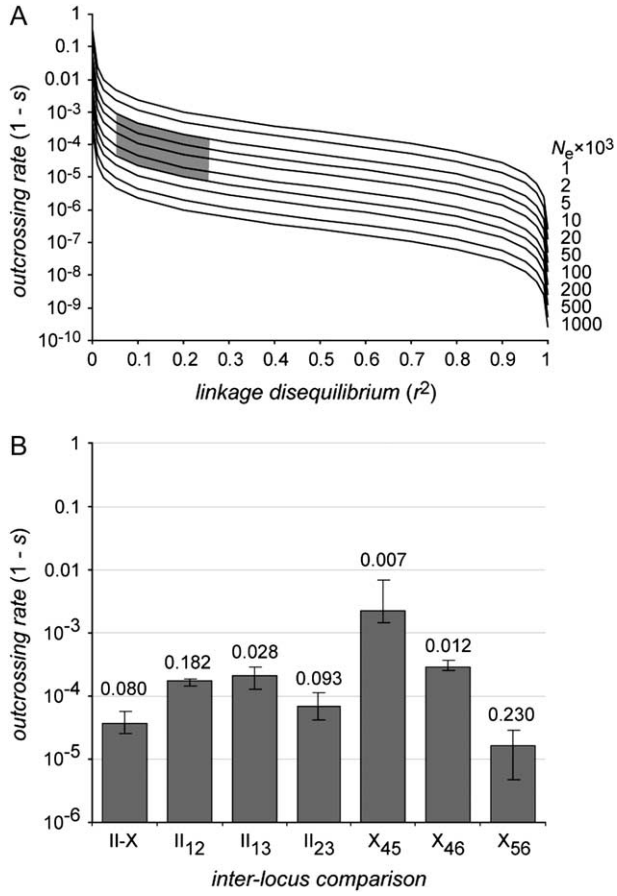

Figure 5.—(A) Relationship between the rate of outcrossing $(1 - s)$ and linkage disequilibrium (r^2) for freely recombining sites ($c = 0.5$) over a range of effective population sizes (N_e). The shaded region corresponds to plausible values in *C. elegans* based on estimates of N_e and pairwise linkage disequilibrium as estimated by r^2. (B) Outcrossing rates inferred from linkage disequilibrium for all pairs of sites between chromosomes (II–X) and between loci i and j within a chromosome (II_{ij}, X_{ij}; 1, Y25C1A.5; 2, ZK430.1; 3, E01G4.6; 4, D1005.1; 5, R160.7; 6, T24D11.1). Error bars indicate the outcrossing rates derived from the 2.5% linkage disequilibrium percentiles of 1000 random "scattered" samples, with values above the bars showing the corresponding mean estimates of linkage disequilibrium (r^2).

of many demes under an island model of migration (Wakeley 1999; Wakeley and Lessard 2003; Lessard and Wakeley 2004), which seem reasonable at least to a first approximation given the values of F_{st} and lack of geographic structure among localities. The resulting mean linkage disequilibrium for interlocus comparisons yields rough estimates of the outcrossing rate $(1 - s)$ in the range of 1.6×10^{-5} to 2.2×10^{-3} (Figure 5), assuming $N_e = 8 \times 10^4$. A similar approach can be used to estimate the outcrossing rate from the population recombination parameter, $\rho_e = 4N_e c(1 - F)/(1 + F)$ (Nordborg 2000; Lessard and Wakeley 2004), such that the outcrossing rate $[1 - s = \rho_e(4N_e c)^{-1}]$ yields values comparable to the r^2 method ($\rho_{II} = 57.5$, $1 - s = 6.8 \times 10^{-4}$; $\rho_X = 2.8$, $1 - s = 2.1 \times 10^{-5}$; given $c_{II} = 0.263$,

$c_X = 0.420$, $N_e = 8 \times 10^4$). Bear in mind that all of these calculations are quite rough and depend on the accuracy of population statistics for which a relatively small number of polymorphic sites from only six loci have been considered. Note, however, that even large deviations in N_e yield low estimated rates of outcrossing (Figure 5A) and that estimates of linkage disequilibrium are higher (and therefore inferred outcrossing rates are lower) when the scattered-sample approach is not taken.

These estimated rates of outcrossing are lower than what one would expect on the basis of the effect of X chromosome nondisjunction producing males and outcrossing ($>10^{-3}$) (Hedgecock 1976; Chasnov and Chow 2002; Stewart and Phillips 2002; Cutter *et al.* 2003a). However, it is important to recognize the difference between the effective outcrossing rate in a genetic sense and the behavioral outcrossing rate, which can involve mating between related partners. Cross-fertilization, even at a high rate, between close relatives ("biparental inbreeding") behaves just like selfing by generating linkage disequilibrium and short times to a common ancestor (Uyenoyama 1986; Nordborg 2000). Barrière and Félix (2005) recently inferred outcrossing rates of $\sim 10^{-5}$ from linkage disequilibrium and $\sim 10^{-2}$ from the frequency of microsatellite heterozygotes of European samples. Also from microsatellite measures of heterozygote frequency, Sivasundar and Hey (2005) suggest outcrossing rates of $\sim$20% in samples from California. Another quantitative estimate of outcrossing from natural isolates is that of Cutter and Payseur (2003), where application of a background selection model to the pattern of SNP density in the genome implies an outcrossing rate of $\sim$1%. The consensus among most of these estimates is that outcrossing is an infrequent, but persistent, phenomenon in *C. elegans*.

Tests of neutrality and population dynamics: Most of the tests of neutrality for the six loci included here suggest an excess of intermediate-frequency alleles at global and regional scales (*i.e.*, Tajima's $D > 0$ or Fu and Li's $D^* > 0$), but not within individual localities. What might cause such skews in the frequency spectrum of alleles? Widespread balancing selection seems unlikely in this case. Heterozygote advantage is not likely, given low heterozygosity due to selfing and the failure to detect heterosis or inbreeding depression in *C. elegans* (Johnson and Wood 1982; Johnson and Hutchinson 1993; Chasnov and Chow 2002). Instead, the trend of positive values of D at global and regional scales (but not among localities) likely reflects mainly population structure, consistent with the observation of moderate F_{st} at the local scale (Tajima 1989; Pannell 2003).

Local population growth tends to cause negative values of Tajima's D, as do selective sweeps and weak background selection (Maynard Smith and Haigh 1974; Tajima 1989; Charlesworth *et al.* 1993). One demographic scenario to test with additional global samples is the possibility of local population growth following a recent postglacial colonization of Europe. In the face of the strong linkage disequilibrium due to selfing, purifying selection against deleterious mutations (background selection) or selective sweeps (genetic hitchhiking) may also be particularly potent forces that could contribute to *C. elegans*' very low diversity, but their effects on D are difficult to discern (Maynard Smith and Haigh 1974; Charlesworth *et al.* 1993; Nordborg 1997). Such selection can reduce genetic variation at linked sites to an extent much greater than the twofold reduction expected from selfing alone (Charlesworth *et al.* 1993).

Negative values of D can also be caused by extinction-recolonization dynamics in a metapopulation, if the extinction rate is sufficiently high (Pannell 2003). For a metapopulation process to influence patterns of polymorphism: (1) the extinction rate must exceed the migration rate and (2) the number of colonists must exceed twice the number of migrants to extant populations under a "migrant-pool" model, generally leading to a combination of high F_{st} and low π_T and π_S (Wade and McCauley 1988; Pannell and Charlesworth 1999; Pannell 2003). Although this qualitative pattern corresponds to our observations at the local scale, nothing is known about extinction rates of *C. elegans* subpopulations and there is little reason to expect the number of colonists to greatly exceed that of migrants. There also is little reason to expect extinction to exceed migration, although this may be testable in the future. Wade and McCauley (1988) argue that when migration and colonization are similar behaviors and occur within the boundaries of the metapopulation, as is likely the case in *C. elegans*, the number of migrants and colonists would be approximately equal. Thus, the patterns of polymorphism in this species may be explained primarily by population isolation caused by inbreeding, coupled with migration between subpopulations at all spatial scales, rather than by turnover of populations.

In many respects, locus Y25C1A.5 is unusual relative to the other loci examined, with significantly positive Tajima's D, high D_{st}, high haplotype diversity, and many indels. Could this locus or a linked locus be generating a signature of local adaptation? The protein product of this gene forms a subunit of the coatomer (COPI) complex associated with vesicle transport (www.wormbase.org). It is expressed in many tissues during both larval and adult development, and application of RNAi affects fertility, adult viability, and osmoregulation (Kamath *et al.* 2003). Our focus on intron sequence precludes the detection of potentially important amino acid polymorphisms, although the protein sequence evolves at an average rate when compared with *C. briggsae* (Table 6). However, many loci are effectively closely linked to this gene, and, given extensive linkage disequilibrium, it may prove difficult to determine the cause of the unusual molecular evolutionary patterns in this region.

Comparison with the partial selfer *A. thaliana*: *A. thaliana* global diversity at silent sites exceeds that of *C. elegans* by about fourfold (Shepard and Purugganan 2003; Nordborg *et al.* 2005; Schmid *et al.* 2005), despite the fact that both are self-fertile hermaphrodites with worldwide human-commensal distributions. The distribution of genetic diversity within and between populations also appears to differ between these two species, with intrapopulation diversity making up a larger portion of the variation in *C. elegans* (Abbott and Gomes 1989; Bergelson *et al.* 1998). Marked differences are also apparent in the variant frequency spectra (*e.g.*, Tajima's *D*). Loci throughout the *A. thaliana* genome in global and regional samples show a skew toward negative values of *D*, indicating an excess of rare variants (Nordborg *et al.* 2005; Schmid *et al.* 2005). Such a pattern at particular loci can be caused by positive selection, but suggests purifying selection and population growth when observed as the background pattern in a genome (Tajima 1989; Nordborg *et al.* 2005; Schmid *et al.* 2005). In contrast, at global and regional scales in *C. elegans*, we find an excess of intermediate-frequency variants (*i.e.*, $\pi > \theta$ and $D > 0$), which is probably a consequence of population structure, because this trend disappears at smaller spatial scales (Tajima 1989; Pannell 2003). In *A. thaliana*, linkage disequilibrium decays over a span of 25–50 kb (Nordborg *et al.* 2005), whereas many pairs of sites on different chromosomes are not in linkage equilibrium in *C. elegans*. These differences in the patterns of variation in the genomes of *C. elegans* and *A. thaliana* suggest that outcrossing may be more prevalent in the plant, but that migration is probably more important in the worm. A provocative ecological hypothesis holds that these differences may be expected if size-dependent dispersal is partially responsible for shaping global patterns of diversity (Finlay 2002).

Implications for *C. elegans* evolution: With a modest effective population size of $\sim 8 \times 10^4$, natural selection will be unable to act efficiently on mutations with very low selection coefficients (s), such as those associated with codon usage bias ($s \sim 10^{-6}$) (Akashi 1999; Maside *et al.* 2004). However, several analyses have detected selection on codon usage bias in the genome of *C. elegans* (Stenico *et al.* 1994; Duret 2000; Marais and Duret 2001; Cutter *et al.* 2003b; Cutter and Ward 2005). If selection is relaxed, codon usage bias decays very slowly over time (Marais *et al.* 2004). Thus, an ancestrally large population that has recently been reduced in size in the lineage leading to *C. elegans*, perhaps due to a recent origin of self-fertilization, could explain the persistence of codon bias. Alternatively, our estimates of the global effective population size based on diversity may be underestimates if much of *C. elegans* diversity has yet to be discovered.

It is also informative to calculate the expected time to the most recent common ancestor of our samples. Under the assumption of no recombination, the expected coalescence time of segregating polymorphisms is 4 N_e generations (although the variance is high). *C. elegans* generation time is ~4 days under laboratory conditions, although a 60-day generation time may be more appropriate if *C. elegans* spend most of their life cycle in the dauer stage (Riddle and Wood 1988; Barrière and Félix 2005). An average 60-day generation time implies that the common ancestor of the French, German, and Scottish nematodes may have lived ~34,000 years ago and that the coalescent for the global CGC samples is ~60,000 years.

If we can assume that the origin of selfing in the *C. elegans* lineage can be traced back to a mutation or series of mutations that swept a single genotype to fixation, then all extant genetic variation will result from subsequent mutation in that original self-fertile genetic background. Consequently, the above calculations of the time to the most recent common ancestor in our sample could provide a lower-bound estimate on how long selfing has persisted in this lineage. However, this lower bound is likely to greatly underestimate the duration of selfing in *C. elegans* for several reasons. First, self-fertilization reduces N_e, and thus speeds up the rate of coalescence, causing coalescent times for extant polymorphism to be much more recent than the origin of selfing itself (Nordborg and Donnelly 1997). Second, selective sweeps will proceed rapidly and remove diversity across the genome, given the levels of selfing and migration, so coalescent times may reflect simply the time to the most recent selective sweep. Third, because no *C. elegans* strains from Asia, Africa, or South America have yet been isolated or analyzed, our current evaluation of *C. elegans* diversity might drastically underestimate global diversity by principally reflecting recent European population processes and emigration to North America and Australia. It remains a challenge to determine how long *C. elegans* has persisted as a self-fertile species, between the large possible temporal bounds of 60 thousand and 100 million years (Stein *et al.* 2003; Kiontke *et al.* 2004).

Discussions with D. Charlesworth were instrumental for the design and analysis of this work. I am also grateful to E. Dolgin and P. Keightley for assistance in field collections; to M. Felix and A. Barriere for instruction in nematode sampling and identification; to M. Blaxter, B. Charlesworth, and K. Dyer for insightful discussion; and to D. Charlesworth, J. Hey, B. Payseur, and an anonymous reviewer for critical comments on the manuscript. M. Felix and the Caenorhabditis Genetics Center kindly provided strains that were used in this study. This work was funded by the National Science Foundation International Research Fellowship Program grant no. 0401897.

LITERATURE CITED

Abbott, R. J., and M. F. Gomes, 1989 Population genetic structure and outcrossing rate of *Arabidopsis thaliana* (L.) Heynh. Heredity **62:** 411–418.

Agapow, P. M., and A. Burt, 2001 Indices of multilocus linkage disequilibrium. Mol. Ecol. Notes **1:** 101–102.

Akashi, H., 1999 Inferring the fitness effects of DNA mutations from polymorphism and divergence data: statistical power to

detect directional selection under stationarity and free recombination. Genetics **151:** 221–238.

Andolfatto, P., 2001 Contrasting patterns of X-linked and autosomal nucleotide variation in *Drosophila melanogaster* and *Drosophila simulans*. Mol. Biol. Evol. **18:** 279–290.

Barrière, A., and M. A. Félix, 2005 High local genetic diversity and low outcrossing rate in *Caenorhabditis elegans* natural populations. Curr. Biol. **15:** 1176–1184.

Bergelson, J., E. Stahl, S. Dudek and M. Kreitman, 1998 Genetic variation within and among populations of *Arabidopsis thaliana*. Genetics **148:** 1311–1323.

Charlesworth, B., M. T. Morgan and D. Charlesworth, 1993 The effect of deleterious mutations on neutral molecular variation. Genetics **134:** 1289–1303.

Charlesworth, B., M. Nordborg and D. Charlesworth, 1997 The effects of local selection, balanced polymorphism and background selection on equilibrium patterns of genetic diversity in subdivided populations. Genet. Res. **70:** 155–174.

Charlesworth, D., 2003 Effects of inbreeding on the genetic diversity of populations. Philos. Trans. R. Soc. Lond. B Biol. Sci. **358:** 1051–1070.

Chasnov, J. R., and K. L. Chow, 2002 Why are there males in the hermaphroditic species *Caenorhabditis elegans*? Genetics **160:** 983–994.

Cutter, A. D., and B. A. Payseur, 2003 Selection at linked sites in the partial selfer *Caenorhabditis elegans*. Mol. Biol. Evol. **20:** 665–673.

Cutter, A. D., and S. Ward, 2005 Sexual and temporal dynamics of molecular evolution in *C. elegans* development. Mol. Biol. Evol. **22:** 178–188.

Cutter, A. D., L. Avilés and S. Ward, 2003a The proximate determinants of sex ratio in *C. elegans* populations. Genet. Res. **81:** 91–102.

Cutter, A. D., B. A. Payseur, T. Salcedo, A. M. Estes, J. M. Good *et al.*, 2003b Molecular correlates of genes exhibiting RNAi phenotypes in *Caenorhabditis elegans*. Genome Res. **13:** 2651–2657.

Delattre, M., and M. A. Félix, 2001 Microevolutionary studies in nematodes: a beginning. BioEssays **23:** 807–819.

Denver, D. R., K. Morris and W. K. Thomas, 2003 Phylogenetics in *Caenorhabditis elegans*: an analysis of divergence and outcrossing. Mol. Biol. Evol. **20:** 393–400.

Denver, D. R., K. Morris, M. Lynch and W. K. Thomas, 2004 High mutation rate and predominance of insertions in the *Caenorhabditis elegans* nuclear genome. Nature **430:** 679–682.

Duret, L., 2000 tRNA gene number and codon usage in the *C. elegans* genome are co-adapted for optimal translation of highly expressed genes. Trends Genet. **16:** 287–289.

Dye, C., and B. G. Williams, 1997 Multigenic drug resistance among inbred malaria parasites. Proc. R. Soc. Lond. Ser. B Biol. Sci. **264:** 61–67.

Fearnhead, P., and P. Donnelly, 2001 Estimating recombination rates from population genetic data. Genetics **159:** 1299–1318.

Finlay, B. J., 2002 Global dispersal of free-living microbial eukaryote species. Science **296:** 1061–1063.

Floyd, R., E. Abebe, A. Papert and M. Blaxter, 2002 Molecular barcodes for soil nematode identification. Mol. Ecol. **11:** 839–850.

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

Graustein, A., J. M. Gaspar, J. R. Walters and M. F. Palopoli, 2002 Levels of DNA polymorphism vary with mating system in the nematode genus Caenorhabditis. Genetics **161:** 99–107.

Haber, M., M. Schungel, A. Putz, S. Muller, B. Hasert *et al.*, 2005 Evolutionary history of *Caenorhabditis elegans* inferred from microsatellites: evidence for spatial and temporal genetic differentiation and the occurrence of outbreeding. Mol. Biol. Evol. **22:** 160–173.

Haubold, B., and R. R. Hudson, 2000 LIAN 3.0: detecting linkage disequilibrium in multilocus data. Bioinformatics **16:** 847–848.

Hedgecock, E. M., 1976 Mating system of *Caenorhabditis elegans*: evolutionary equilibrium between self-fertilization and cross-fertilization in a facultative hermaphrodite Am. Nat. **110:** 1007–1012.

Hein, J., M. H. Schierup and C. Wiuf, 2004 *Gene Genealogies, Variation and Evolution: A Primer in Coalescent Theory*. Oxford University Press, Oxford.

Hill, W. G., and A. Robertson, 1968 Linkage disequilibrium in finite populations. Theor. Appl. Genet. **38:** 226–231.

Hodgkin, J., and T. Doniach, 1997 Natural variation and copulatory plug formation in *Caenorhabditis elegans*. Genetics **146:** 149–164.

Hudson, R. R., and N. L. Kaplan, 1985 Statistical properties of the number of recombination events in the history of a sample of DNA sequences. Genetics **111:** 147–164.

Ingvarsson, P. K., 2002 A metapopulation perspective on genetic diversity and differentiation in partially self-fertilizing plants. Evolution **56:** 2368–2373.

Johnson, T. E., and E. W. Hutchinson, 1993 Absence of strong heterosis for life span and other life-history traits in *Caenorhabditis elegans*. Genetics **134:** 465–474.

Johnson, T. E., and W. B. Wood, 1982 Genetic-analysis of life-span in *Caenorhabditis elegans*. Proc. Natl. Acad. Sci. USA **79:** 6603–6607.

Jovelin, R., B. C. Ajie and P. C. Phillips, 2003 Molecular evolution and quantitative variation for chemosensory behaviour in the nematode genus Caenorhabditis Mol. Ecol. **12:** 1325–1337.

Kamath, R. S., A. G. Fraser, Y. Dong, G. Poulin, R. Durbin *et al.*, 2003 Systematic functional analysis of the *Caenorhabditis elegans* genome using RNAi. Nature **421:** 231–237.

Keightley, P. D., and B. Charlesworth, 2005 Genetic instability of *C. elegans* comes naturally. Trends Genet. **21:** 67–70.

Kiontke, K., N. P. Gavin, Y. Raynes, C. Roehrig, F. Piano *et al.*, 2004 Caenorhabditis phylogeny predicts convergence of hermaphroditism and extensive intron loss. Proc. Natl. Acad. Sci. USA **101:** 9003–9008.

Koch, R., H. G. A. M. van Luenen, M. van der Horst, K. L. Thijssen and R. H. A. Plasterk, 2000 Single nucleotide polymorphisms in wild isolates of *Caenorhabditis elegans*. Genome Res. **10:** 1690–1696.

Kreitman, M., 1983 Nucleotide polymorphism at the alcohol-dehydrogenase locus of *Drosophila melanogaster*. Nature **304:** 412–417.

Lessard, S., and J. Wakeley, 2004 The two-locus ancestral graph in a subdivided population: convergence as the number of demes grows in the island model. J. Math. Biol. **48:** 275–292.

Marais, G., and L. Duret, 2001 Synonymous codon usage, accuracy of translation, and gene length in *Caenorhabditis elegans*. J. Mol. Evol. **52:** 275–280.

Marais, G., D. Mouchiroud and L. Duret, 2001 Does recombination improve selection on codon usage? Lessons from nematode and fly complete genomes. Proc. Natl. Acad. Sci. USA **98:** 5688–5692.

Marais, G., B. Charlesworth and S. I. Wright, 2004 Recombination and base composition: the case of the highly self-fertilizing plant *Arabidopsis thaliana*. Genome Biol. **5:** R45.

Maside, X. L., A. W. S. Lee and B. Charlesworth, 2004 Selection on codon usage in *Drosophila americana*. Curr. Biol. **14:** 150–154.

Maynard Smith, J., and J. Haigh, 1974 Hitch-hiking effect of a favorable gene. Genet. Res. **23:** 23–35.

Mitchell-Olds, T., 2001 *Arabidopsis thaliana* and its wild relatives: a model system for ecology and evolution. Trends Ecol. Evol. **16:** 693–700.

Myers, S. R., and R. C. Griffiths, 2003 Bounds on the minimum number of recombination events in a sample history. Genetics **163:** 375–394.

Nei, M., and W. H. Li, 1979 Mathematical-model for studying genetic-variation in terms of restriction endonucleases. Proc. Natl. Acad. Sci. USA **76:** 5269–5273.

Nordborg, M., 1997 Structured coalescent processes on different timescales. Genetics **146:** 1501–1514.

Nordborg, M., 2000 Linkage disequilibrium, gene trees and selfing: an ancestral recombination graph with partial self-fertilization. Genetics **154:** 923–929.

Nordborg, M., and P. Donnelly, 1997 The coalescent process with selfing. Genetics **146:** 1185–1195.

Nordborg, M., T. T. Hu, Y. Ishino, J. Jhaveri, C. Toomajian *et al.*, 2005 The pattern of polymorphism in *Arabidopsis thaliana*. PLoS Biol. **3:** e196.

Pannell, J. R., 2003 Coalescence in a metapopulation with recurrent local extinction and recolonization. Evolution **57:** 949–961.

Pannell, J. R., and B. Charlesworth, 1999 Neutral genetic diversity in a metapopulation with recurrent local extinction and recolonization. Evolution **53:** 664–676.

Pollak, E., 1987 On the theory of partially inbreeding finite populations. 1. Partial selfing. Genetics **117:** 353–360.

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

Riddle, D. L., and W. B. Wood, 1988 *The Nematode Caenorhabditis elegans*, pp. 393–412. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Rozas, J., J. C. Sanchez-DelBarrio, X. Messeguer and R. Rozas, 2003 DnaSP, DNA polymorphism analyses by the coalescent and other methods. Bioinformatics **19:** 2496–2497.

Schmid, K. J., S. Ramos-Onsins, H. Ringys-Beckstein, B. Weisshaar and T. Mitchell-Olds, 2005 A multilocus sequence survey in *Arabidopsis thaliana* reveals a genome-wide departure from a neutral model of DNA sequence polymorphism. Genetics **169:** 1601–1615.

Shepard, K. A., and M. D. Purugganan, 2003 Molecular population genetics of the Arabidopsis CLAVATA2 region: the genomic scale of variation and selection in a selfing species. Genetics **163:** 1083–1095.

Sivasundar, A., and J. Hey, 2003 Population genetics of *Caenorhabditis elegans*: the paradox of low polymorphism in a widespread species. Genetics **163:** 147–157.

Sivasundar, A., and J. Hey, 2005 Sampling from natural populations using RNAi reveals high outcrossing and population structure in *Caenorhabditis elegans*. Curr. Biol. **15:** 1598–1602.

Stein, L. D., Z. Bao, D. Blasiar, T. Blumenthal, M. R. Brent *et al.*, 2003 The genome sequence of *Caenorhabditis briggsae*: a platform for comparative genomics. PLoS Biol. **1:** 166–192.

Stenico, M., A. T. Lloyd and P. M. Sharp, 1994 Codon usage in *Caenorhabditis elegans*: delineation of translational selection and mutational biases. Nucleic Acids Res. **22:** 2437–2446.

Stewart, A. D., and P. C. Phillips, 2002 Selection and maintenance of androdioecy in *Caenorhabditis elegans*. Genetics **160:** 975–982.

Stewart, M. R., N. L. Clark, G. Merrihew, E. M. Galloway and J. H. Thomas, 2005 High genetic diversity in the chemoreceptor superfamily of *Caenorhabditis elegans*. Genetics **169:** 1985–1996.

Storey, J. D., and R. Tibshirani, 2003 Statistical significance for genomewide studies. Proc. Natl. Acad. Sci. USA **100:** 9440–9445.

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

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

Thomas, W. K., and A. C. Wilson, 1991 Mode and tempo of molecular evolution in the nematode Caenorhabditis: cytochrome oxidase-II and calmodulin sequences. Genetics **128:** 269–279.

Thornton, K., 2003 libsequence: a C++ class library for evolutionary genetic analysis. Bioinformatics **19:** 2325–2327.

Uyenoyama, M. K., 1986 Inbreeding and the cost of meiosis: the evolution of selfing in populations practicing biparental inbreeding. Evolution **40:** 388–404.

Wade, M. J., and D. E. McCauley, 1988 Extinction and recolonization: their effects on the genetic differentiation of local populations. Evolution **42:** 995–1005.

Wakeley, J., 1999 Nonequilibrium migration in human history. Genetics **153:** 1863–1871.

Wakeley, J., and N. Aliacar, 2001 Gene genealogies in a metapopulation. Genetics **159:** 893–905.

Wakeley, J., and S. Lessard, 2003 Theory of the effects of population structure and sampling on patterns of linkage disequilibrium applied to genomic data from humans. Genetics **164:** 1043–1053.

Watterson, G. A., 1975 Number of segregating sites in genetic models without recombination. Theor. Popul. Biol. **7:** 256–276.

Wicks, S. R., R. T. Yeh, W. R. Gish, R. H. Waterston and R. H. A. Plasterk, 2001 Rapid gene mapping in *Caenorhabditis elegans* using a high density polymorphism map. Nat. Genet. **28:** 160–164.

Wright, S. I., I. V. Bi, S. G. Schroeder, M. Yamasaki, J. F. Doebley *et al.*, 2005 The effects of artificial selection of the maize genome. Science **308:** 1310–1314.

Yu, N., Z. Zhao, Y. X. Fu, N. Sambuughin, M. Ramsay *et al.*, 2001 Global patterns of human DNA sequence variation in a 10-kb region on chromosome 1. Mol. Biol. Evol. **18:** 214–222.

Zhang, D. Y., 2000 Resource allocation and the evolution of self-fertilization in plants. Am. Nat. **155:** 187–199.

Zhao, Z., L. Jin, Y. X. Fu, M. Ramsay, T. Jenkins *et al.*, 2000 Worldwide DNA sequence variation in a 10-kilobase noncoding region on human chromosome 22. Proc. Natl. Acad. Sci. USA **97:** 11354–11358.

Communicating editor: M. Nordborg