nature genetics

Chromosome-scale selective sweeps shape *Caenorhabditis elegans* genomic diversity

Erik C Andersen[1,2,7], Justin P Gerke[1,2,6,7], Joshua A Shapiro[1,2,7], Jonathan R Crissman[1,3], Rajarshi Ghosh[1,2], Joshua S Bloom[1,4], Marie-Anne Félix[5] & Leonid Kruglyak[1–3]

The nematode *Caenorhabditis elegans* is central to research in molecular, cell and developmental biology, but nearly all of this research has been conducted on a single strain of *C. elegans*. Little is known about the population genomic and evolutionary history of this species. We characterized *C. elegans* genetic variation using high-throughput selective sequencing of a worldwide collection of 200 wild strains and identified 41,188 SNPs. Notably, *C. elegans* genome variation is dominated by a set of commonly shared haplotypes on four of its six chromosomes, each spanning many megabases. Population genetic modeling showed that this pattern was generated by chromosome-scale selective sweeps that have reduced variation worldwide; at least one of these sweeps probably occurred in the last few hundred years. These sweeps, which we hypothesize to be a result of human activity, have drastically reshaped the global *C. elegans* population in the recent past.

C. elegans is a globally distributed, free-living nematode that colonizes human-associated habitats, including compost heaps and rotting fruit[1]. For the past 40 years, a single laboratory strain of *C. elegans* (N2) has been invaluable to biomedical research as a model for animal development, programmed cell death and RNA interference[2]. Studies of a small number of loci have suggested that *C. elegans* has a small effective population size and low diversity compared to closely related species despite having large local population sizes and global gene flow[3–14]. The factors responsible for this low genetic diversity are unknown. *C. elegans* reproduces primarily by hermaphroditic selfing, but this mating system alone is not sufficient to explain the observed reduction in the diversity of the species[11]. The polymorphism rate between the laboratory strain N2 and the *C. elegans* wild isolate CB4856 correlates with the recombination rate, suggesting that background selection against deleterious mutations also reduces diversity[5,15,16]. However, CB4856 is genetically isolated from the rest of the *C. elegans* population[17], and analyses based on the divergence of CB4856 alone are subject to substantial ascertainment bias and may not fully capture evolutionary processes relevant to the global population. To obtain a more complete description of *C. elegans* diversity, we sequenced thousands of genome fragments from a globally distributed collection of 200 wild isolates. Our results show that recent strong sweeps of positive selection have drastically reduced chromosome-wide diversity in this species.

RESULTS

C. elegans genome diversity and strain relationships

We studied 200 wild strains of *C. elegans* from 58 collection locations on six continents (**Fig. 1** and **Supplementary Table 1**). These strains cover virtually every known collection location for this species, providing the most comprehensive set of *C. elegans* strains assembled to date. The samples were isolated from a variety of sources, including rotting fruits, compost, mushroom farms, soil and snails. To characterize genomic variation among these strains, we examined restriction-site–associated DNA (RAD)[18] covering 8% of the 100-Mb genome. We sequenced 91 bp on both sides of each *EcoR*I restriction site, yielding, on average, a pair of RAD tags every 2.1 kb. We achieved a median coverage of 27 reads per tag per strain, allowing for SNP identification with a false discovery rate (FDR) of less than 0.6% (Online Methods). Across all strains, we identified 41,188 SNPs in 8 Mb of sequence (with an average of 5.1 SNPs per kb).

C. elegans reproduces primarily as a selfing hermaphrodite, which can lead to clonal expansions of a single genotype. For this reason, we expected to find identical strains among samples collected from nearby locations. To find instances of this, we examined the number and distribution of discordant genotype calls across all pairwise strain comparisons. We considered pairs with fewer SNPs than the expected number of false positives given our FDR (250 SNPs) to be clonal, with the exception of the pair ED3046 and ED3049, for which we found the SNPs to be clustered in a small region on chromosome II. Of the 200 sampled strains, 47 had unique haplotypes. The remaining 153 strains grouped into 50 sets of near-identical strains (**Supplementary Table 1**). Most of these sets were from a single isolation or from separate samples in close proximity, probably representing strains sampled from a single clonal expansion. However, two sets spanned different continents: set AB2 and CX11258 from Australia and the United States, respectively, and set JU1171 and MY23

[1]Lewis-Sigler Institute for Integrative Genomics, Princeton University, Princeton, New Jersey, USA. [2]Department of Ecology and Evolutionary Biology, Princeton University, Princeton, New Jersey, USA. [3]Howard Hughes Medical Institute, Princeton University, Princeton, New Jersey, USA. [4]Department of Molecular Biology, Princeton University, Princeton, New Jersey, USA. [5]Institute of Biology of the Ecole Normale Supérieure, Paris, France. [6]Present address: Pioneer Hi-Bred International, A DuPont Business, Johnston, Iowa, USA. [7]These authors contributed equally to this work. Correspondence should be addressed to L.K. (leonid@genomics.princeton.edu).

Received 7 July 2011; accepted 1 December 2011; published online 29 January 2012; doi:10.1038/ng.1050

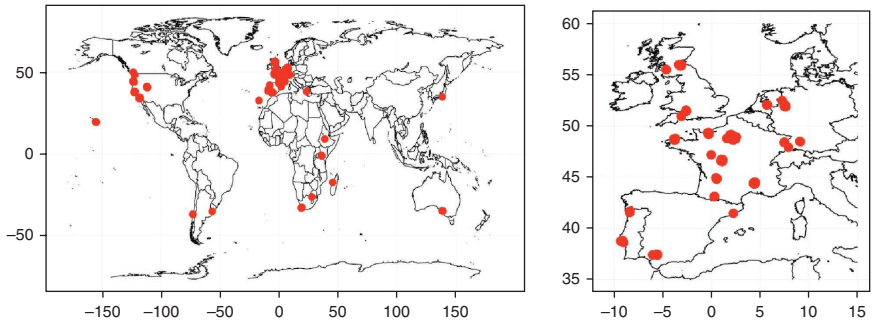

Figure 1 Global sampling locations of *C. elegans* strains. The isolation locations of the wild strains sequenced in this study are shown as red circles on the world map. At right is a map of the more densely sampled western Europe.

from Chile and Germany, respectively. It is possible that these pairs of strains are the result of recent long-range migrations. However, given previous evidence of strain confusion with wild strains isolated before adequate record keeping was practiced[17,19] and our own results, we conservatively analyzed only one strain from each of these sets. The set of 97 distinct genome-wide haplotypes, referred to as 'isotypes' in the subsequent analyses, comprises one isolate from each of the 50 near-identical sets and 47 unique isolates (**Supplementary Table 1**). Phylogenetic clustering of the isotypes revealed little to no grouping by isolation environment or by country of origin (**Fig. 2** and **Supplementary Fig. 1**) but did identify four highly diverged isotypes: CB4856, DL238, JU775 and QX1211. We identified an average of 3,613 SNPs per isotype that differed from the reference strain N2, but these four diverged isotypes had an average of such 9,141 SNPs. In particular, 18% of the variants in the full SNP set are found only in QX1211, which was isolated in San Francisco.

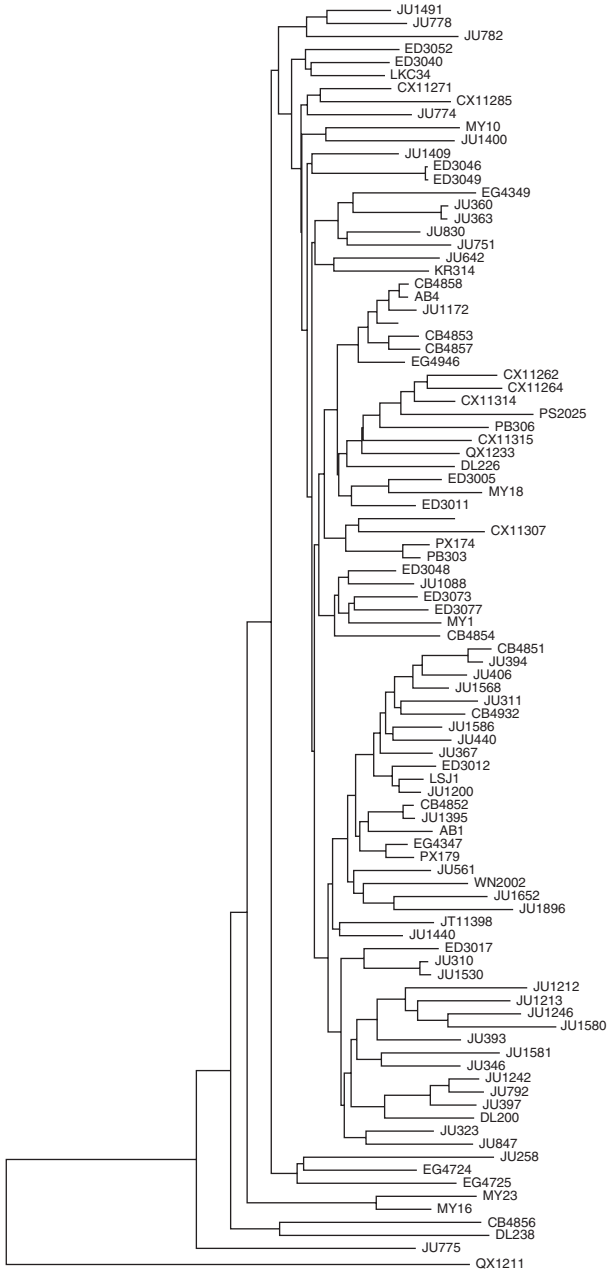

Figure 2 Neighbor-joining tree of 97 *C. elegans* isotypes. The neighbor-joining tree was constructed using 40,857 polymorphisms in the set of 97 isotypes and was pseudo-rooted to the QX1211 strain for visualization reasons. The branch lengths are proportional to the number of polymorphisms that differentiate each pair.

Linkage disequilibrium and population structure

Among the isotypes, we found several large blocks of strong linkage disequilibrium (LD) ($r^2 > 0.6$) extending several Mb within chromosomes (**Supplementary Fig. 2**). We also found substantial LD to exist between chromosomes, with r^2 values often above 0.2. The population recombination rate ($4Nr$) on each chromosome, estimated by the composite likelihood[20], ranged from 90 to 185, suggesting an outcrossing rate of between 1/100 and 1/1,000 per generation, depending on the estimate of the effective population size. To test for population subdivision, we used STRUCTURE[21,22] and found statistical support for the existence of only one worldwide population (**Supplementary Fig. 3**). These results suggest that the observed LD is caused mainly by selfing rather than by separation into distinct subpopulations. A principal component analysis (PCA) identified five major axes that explain 29.7% of the genetic variation (**Supplementary Fig. 3**). These axes reveal some geographic structure but do not clearly separate isotypes into distinct subpopulations. There was a weak correlation between geographic distance and genetic relatedness at the local scale (defined as less than 700 km; **Supplementary Fig. 4**), but we found no correlation at larger distances, which is in agreement with previous analyses[3,6,13].

Despite the extensive observed LD, previous results showed the feasibility of genome-wide association analyses in *C. elegans* by mapping two qualitative traits, hybrid incompatibility and copulatory plugging, using SNPs that differed between N2 and CB4856 (ref. 17). Because the causal variants for these traits are known and have a near perfect genotype-phenotype correspondence, we genotyped these variants as proxies for the traits and showed that our set of SNPs can be used to map the variants to the correct genomic regions (**Supplementary Fig. 5**). We also applied association mapping to two quantitative traits (Online Methods and **Supplementary Fig. 5**). Resistance to abamectin, an anthelmintic compound produced by the common soil bacterium *Streptomyces avermitilis*[23], was significantly associated with a 28-kb haplotype on chromosome V

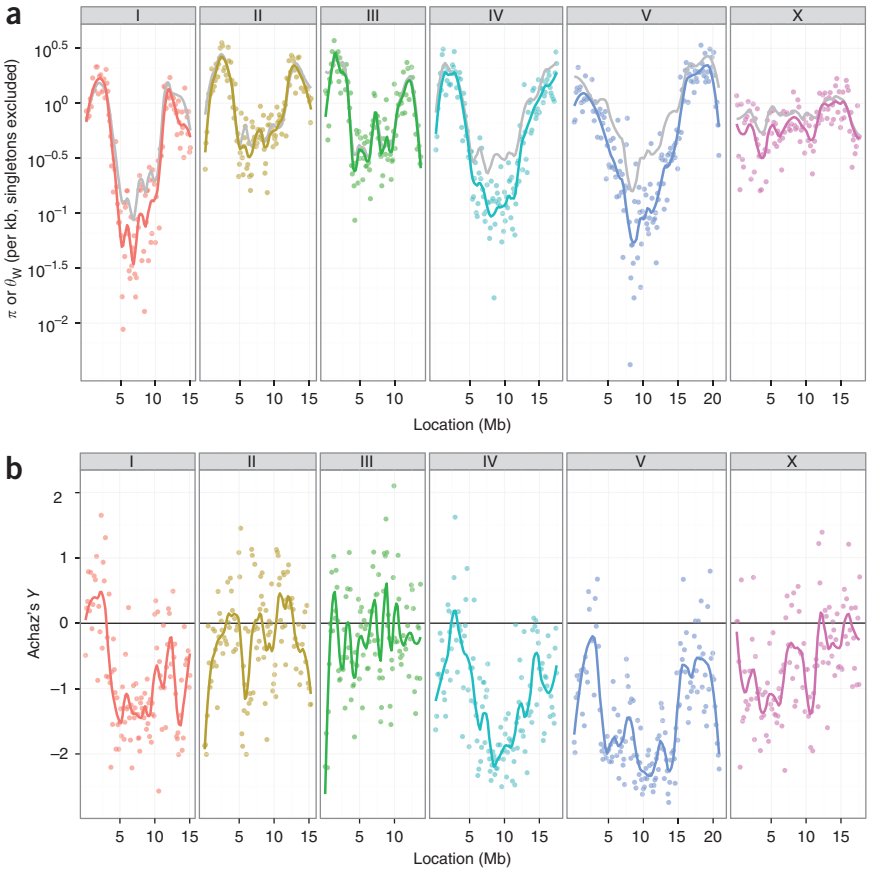

Figure 3 Chromosomal patterns of sequence polymorphism. (**a**) Two estimates of population polymorphism rate, π (colored points and lines) and θ_W (gray lines), are shown for each chromosome. Each point represents a non-overlapping window of 110 RAD tags (approximately 10 kb of sequence). The lines show a locally weighted polynomial regression. (**b**) Achaz's *Y* values, a measure of deviation from the neutral allele frequency spectrum, calculated over the same windows using local polynomial regression. Negative values indicate an excess of rare alleles.

($P = 2.98 \times 10^{-7}$), and aversion to the human pathogen *Pseudomonas aeruginosa* mapped to a 45-kb interval on chromosome IV ($P = 7.45 \times 10^{-9}$). Because geographic structure might be observable using association analyses, we mapped the latitude at which a strain was isolated and found a significant locus in the center of chromosome II ($P = 4.04 \times 10^{-6}$). This association could reflect subtle population structure, or it might implicate this region in an unknown ecological niche preference, such as temperature.

Using these 97 *C. elegans* isotypes, association analyses will probably discover only those alleles that have large phenotypic effects. Additionally, the chromosomal location of the causal variant limits the resolution of the mapping. Extensive LD in the center of a chromosome results in haplotype blocks that are over a Mb in size, as shown here for the traits copulatory plugging and latitude. In contrast, causal variants on the more freely recombining chromosome arms can be localized to haplotype blocks smaller than 50 kb, as is shown here for hybrid incompatibility, abamectin resistance and *P. aeruginosa* avoidance. In this regard, it is worth noting that functional variants in *C. elegans* are more likely to be located on chromosome arms because of the correlation between rates of recombination and polymorphism[15].

Genetic variation and chromosome-wide haplotype sharing

Despite a global distribution, with local populations probably containing millions of individuals[6], our results confirm that genetic variation in *C. elegans* is low. Our genome-wide coverage shows that the amount of diversity in this species varies across genomic regions (**Fig. 3**), as was suggested by previous results derived from a small number of loci[5,24]. The estimated population mutation rate of *C. elegans* (θ_W; Online Methods) varies over two orders of magnitude, from greater than 3.5×10^{-3} per bp on some chromosome arms to a minimum of 2.5×10^{-5} per bp in the centers, averaging out at 8.3×10^{-4} per bp. The amount of polymorphism correlates with the recombination rate on all autosomes[17]; diversity is lower in the low-recombination chromosome centers and higher on the more freely recombining

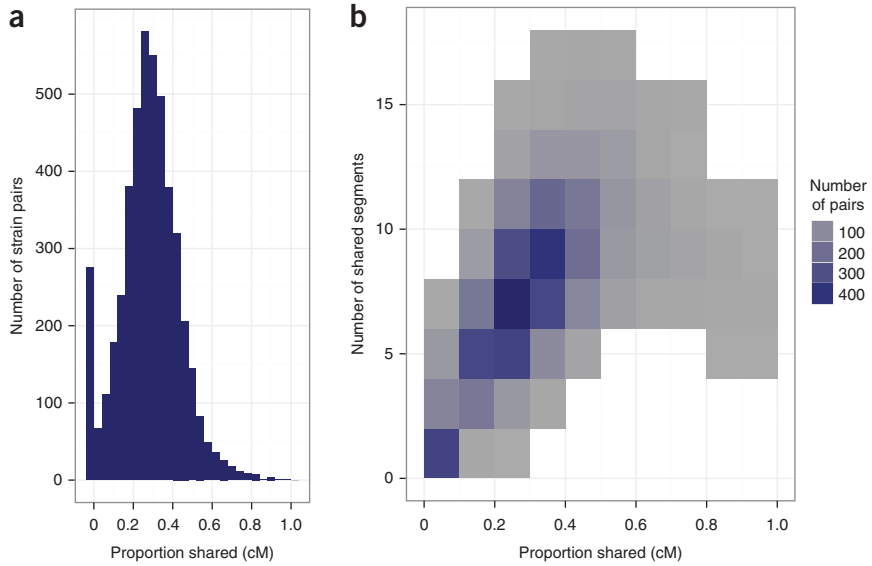

Figure 4 Extensive sharing of large blocks of near-identical haplotypes. (**a**) Proportion of the genetic map shared (as determined by GERMLINE[26]) for every pairwise comparison of the 97 isotypes is shown as a histogram. Notably, every pairwise comparison containing one of the most diverged isotypes (CB4856, DL238 or QX1211) showed little to no sharing. By contrast, the average sharing between a pair was one third of the genetic map. (**b**) Two-dimensional density plot of the number of shared segments and the proportion of the genetic map shared show that most isotype pairs share about one third of the genome in six to ten segments, indicating that the shared segments are large.

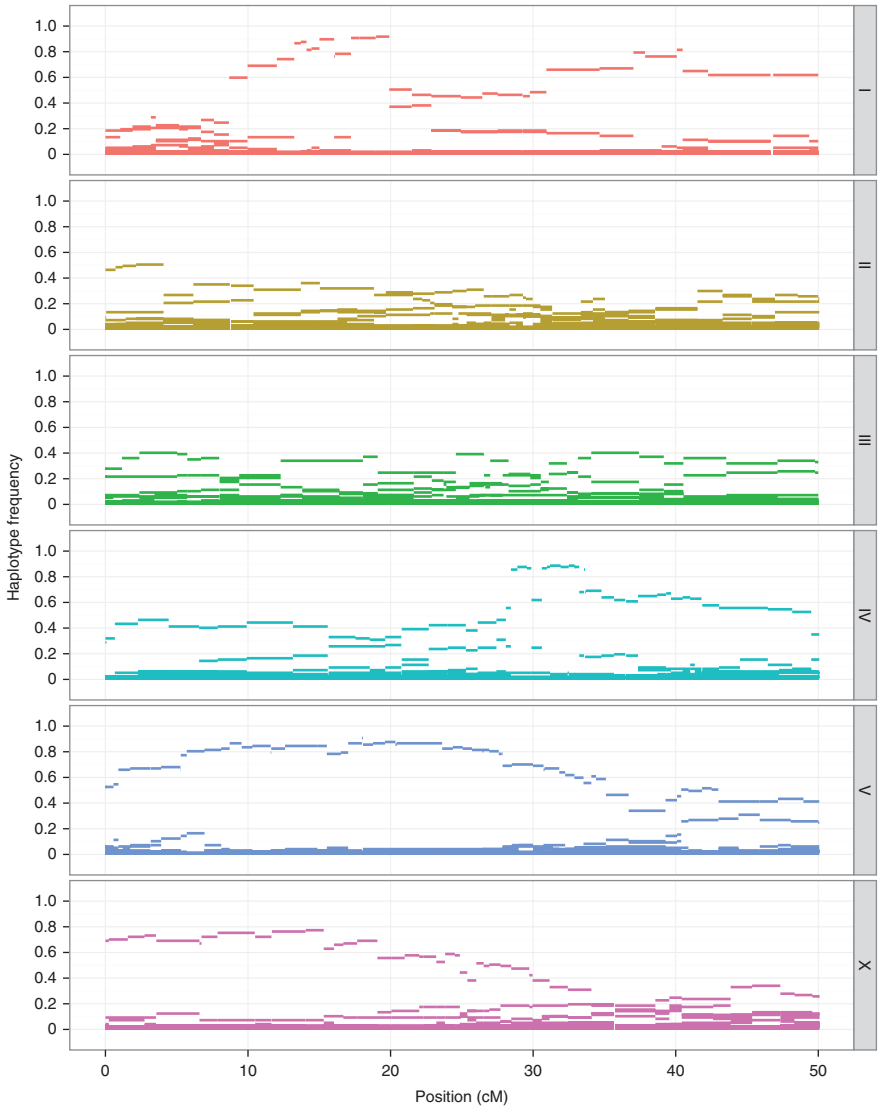

Figure 5 High-frequency extended haplotypes on chromosomes I, IV, V and X. The chromosomal region (in cM) covered by each haplotype block is shown as a bar along the *x* axis, with the haplotype frequency indicated by the height on the *y* axis.

arms (**Fig. 3** and **Supplementary Fig. 6**). On the X chromosome, this pattern is much weaker, and the amount of polymorphism is fairly constant across the entire length of this chromosome ($\theta_W \sim 8.5 \times 10^{-4}$), which corresponds to the more uniform recombination rate of this chromosome[17]. The correlation between rates of polymorphism and recombination is consistent with previous results implicating background selection as a major force shaping patterns of *C. elegans* diversity[5,15]. Variation in pairwise diversity (π) follows the same general pattern as that of θ_W, but there is a larger reduction in π than in θ_W in the centers of chromosomes I, IV and V. This difference results in extremely negative values of Achaz's *Y* (an analog of Tajima's *D*; Online Methods) and indicates an excess of low-frequency polymorphism relative to that in the neutral expectation (**Fig. 3** and **Supplementary Fig. 6**). The left arm of the X chromosome also has an excess of rare variants, but unlike on chromosomes I, IV and V, this region does not have a low recombination rate.

The genome of the wild strain CB4858 seems to contain large haplotypes that are shared with the reference strain N2 (ref. 25), indicating a recent common ancestry between these two strains. To identify whether additional such relationships exist among the 97 isotypes, we used the program GERMLINE[26] to search each pair for segments of at least 2 cM or Mb with no more than two SNP differences, which we defined as 'shared' segments (Online Methods). Notably, we found extensive sharing of large haplotypes among the majority of the isotypes, suggesting a recent common ancestry (**Fig. 4**). The average pair shared roughly one third of the genome identical by descent when measured on either the genetic (median of 28%) or the physical map (median of 33%). The median block size of the shared segments was roughly a fifth of a chromosome (2.5 Mb). Some blocks spanned more than a third of a chromosome, indicating that very few generations of outcrossing have occurred since the most recent common ancestor. Most notably, the patterns of sharing were unevenly distributed across the genome; 70–90% of isotypes shared segments that spanned several Mb on chromosomes I, IV, V and X (**Fig. 5**), but we did not observe such sharing on chromosomes II and III (**Supplementary Fig. 7**). In particular, chromosome V had one common haplotype that spanned the majority of its length.

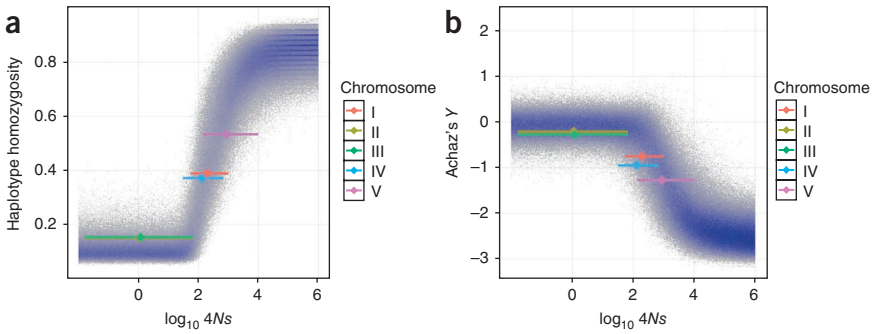

Figure 6 Modeled effects of selection. Results from 10^6 coalescent simulations of chromosomes with a single positively selected site in the center of the chromosome (Online Methods) are plotted. Regions with a high density of points are indicated in blue. (**a**,**b**) Haplotype homozygosity (**a**) and Achaz's *Y* (**b**) for the entire chromosome are plotted against the simulated selection coefficient 4*Ns*. The values observed in our experimental data for each chromosome are indicated by the vertical positions of the colored diamonds. The location of each diamond on the *x* axis is the median 4*Ns* value as estimated from the simulated data, with the length of the bar showing the 90% credible interval.

These regions of high haplotype homozygosity corresponded to the regions with an excess of rare SNPs noted above. Notably, we found the common haplotypes for chromosomes I, IV and V on all six sampled continents; the chromosome X common haplotype was present on five continents.

Recent strong selective sweeps

The combination of high haplotype homozygosity extending over large regions and an excess of rare variants is expected after a strong selective sweep, especially when the recombination rate in a population is low[27]. To estimate the population and selection parameters required to generate the patterns observed here, we performed coalescent simulations of entire chromosomes over a range of demographic models, including single and multiple populations with varying migration rates and population sizes. All models incorporated the effects of background selection and recombination on chromosomal diversity patterns. Demographic forces and background selection are expected to affect the patterns of variation on all chromosomes equally, resulting in a single best-fitting model. Contrary to this expectation, the patterns of variation on chromosomes II and III were markedly different from the patterns on chromosomes I, IV and V. Although the patterns of polymorphism on chromosomes II and III were compatible with models that did not include positive selection, fitting both the excess of rare variants and the high haplotype homozygosity observed for chromosomes I, IV and V required incorporating positive selection (Online Methods and **Fig. 6**; we did not test the X chromosome). Our estimates of the population selection parameter $4Ns$ for these chromosomes ranged from a minimum of 100 to a maximum of 500. For an effective population size of between 10,000 and 25,000 individuals, these values of $4Ns$ correspond to a selective advantage in the range of 0.1–1.3% per generation.

To estimate the timing of these selective sweeps, we focused on the largest and most highly shared segment, which is found on chromosome V and is shared by 84 of the 97 isotypes. Using coalescent simulations with two different models of population growth (**Supplementary Fig. 8**), we estimated that this haplotype arose between 600 and 1,250 generations ago (90% credible interval). In the laboratory, *C. elegans* can go through 100 generations per year, but the average generation time in nature is probably much longer[4]. If we assume a conservative estimate of six generations per year[28], the common haplotype on chromosome V probably expanded to its current frequency in the past 100–200 years. A lower bound on this estimate is provided by the strain CB4851, which was isolated before the year 1949 and carries the selected haplotypes on chromosomes I, V and X, making it probable that those sweeps began no less than 60 years ago. Even if the effective population size and generation time differ by an order of magnitude from our estimated values, the selective sweep still would have occurred in historical times.

DISCUSSION

We report the most comprehensive survey of *C. elegans* diversity to date. Our results indicate that polymorphism rates in this species are correlated with recombination rates, that LD extends over long distances and often occurs between loci on different chromosomes and that there is little detectable subdivision of the global population. Notably, we found extensive sharing of large haplotypes on a subset of chromosomes, accompanied by a paucity of common variation in these regions. The shared haplotypes are distributed geographically throughout the world (**Supplementary Fig. 9**). These observations can only be explained by one or more strong recent global sweeps that were driven by positive selection, a scenario previously considered unlikely in *C. elegans* specifically and *Caenorhabditis* in general[5,29]. Only QX1211 (a strain recently isolated in California), CB4856 and DL238 (a strain isolated in Hawaii) do not share any large haplotypes with the rest of the isotypes. These observations suggest that Hawaii and the Pacific Rim may be fruitful ground for discovery of additional highly diverged isolates. Focused searches in these locations, as well as in other poorly sampled parts of the globe, may yield strains that represent the broader *C. elegans* diversity that existed before the selective sweeps that homogenized much of the global population of this species.

Identification of the beneficial alleles that swept through the *C. elegans* population will be challenging. A selective advantage of 0.1–1% per generation is sufficient to drive a rapid selective sweep in nature, but phenotypic differences of that magnitude are difficult to reliably detect in the laboratory. Within each swept region, there are hundreds of genes with potential effects on fitness, and we can also only speculate about the selective forces that drove these sweeps. We know little about *C. elegans* ecology[1], and it is possible that selection occurred for adaptation to a specific, currently unknown microenvironment.

Positive selection has reduced *C. elegans* genetic variation on a scale not previously observed in multicellular organisms. The rapid global spread of the selected haplotypes during the last few centuries suggests the possibility that the selected alleles may be related to the association between *C. elegans* and human activity. Long-range human travel and transportation of agricultural products during this time period probably contributed to the global spread of the selected haplotypes. Loci that aid human-assisted dispersal and/or confer fitness advantages in human-associated habitats may have driven the observed sweeps. Whether the sweeps resulted in global replacement of endemic populations or *de novo* colonization of new environments by *C. elegans* is unclear. The evolution of the parasitic protozoan *Toxoplasma gondii* provides a notable parallel. Like *C. elegans, T. gondii* is a small human-associated eukaryote with a selfing life stage. A chromosome-wide selective sweep of a single haplotype that originated around 10,000 years ago spread throughout the world population of *T. gondii*[30], suggesting that the dramatic changes in human civilization during this period (such as animal domestication) could have had a role in the rapid evolution of a new lineage in this species. Recent drastic alterations of global environments by humans, and the creation of human-associated niches, may have made such selective sweeps a common feature of the genomes of many species.

URLs. More descriptive methods and additional information are available on the Lewis-Sigler Institute website at http://genomics-pubs.princeton.edu/Diversity/. We used RepeatMasker for some of our analyses, available at http://www.repeatmasker.org/.

METHODS

Methods and any associated references are available in the online version of the paper at http://www.nature.com/naturegenetics/.

Accession codes. All reads are archived in the NCBI Sequence Read Archive under the accession number 047839. SNP calls for all strains (including singleton SNPs) have been uploaded to dbSNP and available under the submitter handle 'KRUGLYAK'.

Note: Supplementary information is available on the Nature Genetics website.

ACKNOWLEDGMENTS

We thank all the contributors who provided strains to make this study possible, including M. Ailion (University of Washington), A. Cutter (University of Toronto), D. Denver (Oregon State University), S. Estes (Portland State University), J. Hodgkin (University of Oxford), J. Kammenga (Wageningen University), P. McGrath (Rockefeller University), M. Rockman (New York University) and

the *Caenorhabditis* Genetics Center. We also thank the Waksman Genomics Core Facility and the Lewis-Sigler Institute Microarray Facility for the Illumina sequencing. We also thank P. Andolfatto, M. Rockman, H. Seidel, R. Tanny and the members of the Kruglyak laboratory for helpful discussions and comments on the manuscript. This work was supported by a US National Institutes of Health (NIH) Ruth L. Kirschstein National Research Service Award (E.C.A.), the Merck Fellowship of the Life Science Research Foundation (J.P.G.), a National Science Foundation graduate research fellowship (J.S.B.), a James S. McDonnell Foundation Centennial Fellowship, the Howard Hughes Medical Institute and NIH grants R01-HG004321, R37-MH59520 (L.K.) and P50-GM071508 to the Center for Quantitative Biology at the Lewis-Sigler Institute of Princeton University.

AUTHOR CONTRIBUTIONS

E.C.A. and J.P.G. conceived of and carried out the experiments and data analyses. J.A.S. conceived of and carried out the data processing, simulations and analysis. J.R.C. assisted with DNA preparations. R.G. performed the abamectin sensitivity assays. J.S.B. analyzed and plotted the LD data. M.-A.F. provided the unpublished wild strains. L.K. supervised all aspects of the study. The manuscript was written by E.C.A., J.P.G., J.A.S. and L.K.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Published online at http://www.nature.com/naturegenetics/.
Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Félix, M.A. & Braendle, C. The natural history of *Caenorhabditis elegans*. *Curr. Biol.* **20**, R965–R969 (2010).
2. Riddle, D.L., Blumenthal, T., Meyer, B.J. & Priess, J.R. *C. Elegans II*, xvii, 1222 (Cold Spring Harbor Laboratory Press, Plainview, New York, USA, 1997).
3. Barrière, A. & Felix, M.A. High local genetic diversity and low outcrossing rate in *Caenorhabditis elegans* natural populations. *Curr. Biol.* **15**, 1176–1184 (2005).
4. Barrière, A. & Felix, M.A. Temporal dynamics and linkage disequilibrium in natural *Caenorhabditis elegans* populations. *Genetics* **176**, 999–1011 (2007).
5. Cutter, A.D. & Payseur, B.A. Selection at linked sites in the partial selfer *Caenorhabditis elegans*. *Mol. Biol. Evol.* **20**, 665–673 (2003).
6. Cutter, A.D. Nucleotide polymorphism and linkage disequilibrium in wild populations of the partial selfer *Caenorhabditis elegans*. *Genetics* **172**, 171–184 (2006).
7. Cutter, A.D., Baird, S.E. & Charlesworth, D. High nucleotide polymorphism and rapid decay of linkage disequilibrium in wild populations of *Caenorhabditis remanei*. *Genetics* **174**, 901–913 (2006).
8. Cutter, A.D., Felix, M.A., Barriere, A. & Charlesworth, D. Patterns of nucleotide polymorphism distinguish temperate and tropical wild isolates of *Caenorhabditis briggsae*. *Genetics* **173**, 2021–2031 (2006).
9. Denver, D.R., Morris, K. & Thomas, W.K. Phylogenetics in *Caenorhabditis elegans*: an analysis of divergence and outcrossing. *Mol. Biol. Evol.* **20**, 393–400 (2003).
10. Dolgin, E.S., Felix, M.A. & Cutter, A.D. Hakuna nematoda: genetic and phenotypic diversity in African isolates of *Caenorhabditis elegans* and *C. briggsae*. *Heredity* **100**, 304–315 (2008).
11. Graustein, A., Gaspar, J.M., Walters, J.R. & Palopoli, M.F. Levels of DNA polymorphism vary with mating system in the nematode genus *Caenorhabditis*. *Genetics* **161**, 99–107 (2002).
12. Haber, M. *et al.* 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 (2005).
13. Sivasundar, A. & Hey, J. Population genetics of *Caenorhabditis elegans*: the paradox of low polymorphism in a widespread species. *Genetics* **163**, 147–157 (2003).
14. Sivasundar, A. & Hey, J. Sampling from natural populations with RNAi reveals high outcrossing and population structure in *Caenorhabditis elegans*. *Curr. Biol.* **15**, 1598–1602 (2005).
15. Rockman, M.V., Skrovanek, S.S. & Kruglyak, L. Selection at linked sites shapes heritable phenotypic variation in *C. elegans*. *Science* **330**, 372–376 (2010).
16. Swan, K.A. *et al.* High-throughput gene mapping in *Caenorhabditis elegans*. *Genome Res.* **12**, 1100–1105 (2002).
17. Rockman, M.V. & Kruglyak, L. Recombinational landscape and population genomics of *Caenorhabditis elegans*. *PLoS Genet.* **5**, e1000419 (2009).
18. Baird, N.A. *et al.* Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS ONE* **3**, e3376 (2008).
19. McGrath, P.T. *et al.* Quantitative mapping of a digenic behavioral trait implicates globin variation in *C. elegans* sensory behaviors. *Neuron* **61**, 692–699 (2009).
20. Hudson, R.R. Two-locus sampling distributions and their application. *Genetics* **159**, 1805–1817 (2001).
21. Falush, D., Stephens, M. & Pritchard, J.K. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics* **164**, 1567–1587 (2003).
22. Pritchard, J.K., Stephens, M. & Donnelly, P. Inference of population structure using multilocus genotype data. *Genetics* **155**, 945–959 (2000).
23. Cully, D.F. *et al.* Cloning of an avermectin-sensitive glutamate-gated chloride channel from *Caenorhabditis elegans*. *Nature* **371**, 707–711 (1994).
24. Koch, R., van Luenen, H.G., van der Horst, M., Thijssen, K.L. & Plasterk, R.H. Single nucleotide polymorphisms in wild isolates of *Caenorhabditis elegans*. *Genome Res.* **10**, 1690–1696 (2000).
25. Hillier, L.W. *et al.* Whole-genome sequencing and variant discovery in *C. elegans*. *Nat. Methods* **5**, 183–188 (2008).
26. Gusev, A. *et al.* Whole population, genome-wide mapping of hidden relatedness. *Genome Res.* **19**, 318–326 (2009).
27. Sabeti, P.C. *et al.* Detecting recent positive selection in the human genome from haplotype structure. *Nature* **419**, 832–837 (2002).
28. Cutter, A.D. Divergence times in *Caenorhabditis* and *Drosophila* inferred from direct estimates of the neutral mutation rate. *Mol. Biol. Evol.* **25**, 778–786 (2008).
29. Phillips, P.C. One perfect worm. *Trends Genet.* **22**, 405–407 (2006).
30. Khan, A., Taylor, S., Ajioka, J.W., Rosenthal, B.M. & Sibley, L.D. Selection at a single locus leads to widespread expansion of *Toxoplasma gondii* lineages that are virulent in mice. *PLoS Genet.* **5**, e1000404 (2009).

ONLINE METHODS

Strains. *C. elegans* were cultured with the bacterial strain OP50 on modified nematode growth medium (NGM)[31] containing 1% agar and 0.7% agarose to prevent burrowing of wild *C. elegans* isolates. Strain information is listed in **Supplementary Table 1**. These strains represent at least one clone from every known isolation location. For locations with more than one strain, we chose strains isolated from different substrates.

A sequence analysis identified strains for which the true identity is suspect. The *Caenorhabditis* Genetics Center (CGC) versions of strains CB4855 and CB4858 were found to be identical by a sequence comparison, even though the strains are reportedly from different isolation locations. The versions of CB4855 and CB4858 from J. Hodgkin are different from each other and from their respective CGC versions but were not used in our analyses. Instead, we treated these two samples as one strain from an unknown location. JU1615 and JU1616 from Melbourne, Australia are likely N2 contaminants, as determined by sequence and behavioral assays; these strains were excluded from our analyses. PX174 and RC301 were found to be identical, despite reported isolations from the United States and Germany, respectively. PX174 was probably a mislabeled RC301 stock and so was excluded from our analyses. JU813 and ED3054 were found to be *Caenorhabditis briggsae* by sequence[32], and mating tests so were not included in any of the analyses.

We also sequenced the following strains, but their sequence or mapping qualities were not high enough to include them in the downstream analyses: CB4855 (the J. Hodgkin version), CX11254 and WN2001.

RAD marker library construction and sequence determination. We isolated genomic DNA by washing nearly starved animals from five 10-cm NGM plates into 15-ml conical tubes and settling them by gravity for 1 h. Genomic DNA was prepared using the DNeasy Blood and Tissue Kit (QIAGEN). Seventeen RAD marker libraries were constructed using Floragenex. Nine additional libraries were constructed using a protocol adapted from previous work[18]. Illumina Genome Analyzer IIx protocols were used for sequencing at 101 cycles.

SNP determination. Each sequence read was entered into a custom mySQL database. Reads were grouped by strain, checked for the presence of a complete *EcoR*I cut sequence and mapped to the WS210 version of the N2 genome using Burrows-Wheeler Aligner (BWA)[33]. Loci with sequence from only a single strain or with fewer than five reads per strain were excluded, as were locations that were less than 100 bp from another cut site. Reads that passed these filters were exported to SAMtools[34] for SNP identification using the 'pileup' command. Called SNPs not in repetitive regions (defined using RepeatMasker) were imported into R, where all subsequent analyses were performed.

We determined an optimal SNP calling strategy using a comparison of libraries generated from different biological replicates of the same strain. Given a quality threshold, sites that differed between replicate libraries were considered errors, and sites that corresponded in both libraries but that differed from the reference were counted as true SNPs. SNPs were called at a phred score threshold of 60, requiring that at least one of each allele have a score ≥120. This approach provided the best balance between a low FDR (~0.6%) and the power to identify true SNPs, yielding 41,188 SNPs. Any genotype call with a score below 60 was considered missing data, and sites missing in more than 25 strains were removed. We imputed missing genotype calls using NPUTE[35]. Both the imputed and un-imputed datasets were condensed from the 200 strains into the 97 isotypes. This reduction eliminated a small number of segregating polymorphisms, resulting in 40,857 SNPs. This SNP set was used for all analyses, except for the association analysis and the detection of population structure.

For the analysis using STRUCTURE[21,22] and for PCA, we constructed a more stringent SNP set. As above, SNPs present in at least one isotype with a quality score of at least 120 were then called in all other isotypes with a quality score threshold of 100. SNPs that were missing or low quality in more than six isotypes were removed. The more stringent cutoff resulted in a set of 6,089 high-quality SNPs. The remaining missing calls were imputed using the program NPUTE[35].

To construct a SNP set for association mapping, we used the 6,089-SNP set but raised the minor allele cutoff to 10 out of 97 isotypes, yielding 4,690 high-quality common SNPs. Missing calls were imputed with NPUTE[35].

Determination of population structure. For the STRUCTURE[21,22] analysis and PCA, we pruned the 6,089-SNP set in sliding 25-marker windows at 5-marker steps, pruning pairs with $r^2 > 0.3$ using PLINK[36]. This reduced the data to 757 SNPs. The results from STRUCTURE[21,22] and PCA were similar at different levels of pruning or missing data thresholds. We used EIGENSOFT[37] for PCA and evaluated significance using Tracy-Widom statistics. Running EIGENSOFT[37] with the 'missingmode: YES' option confirmed that the observed patterns were not caused by structure in the missing data.

Association mapping. We used EMMA[38] for all association analyses with the default kinship matrix. We ignored significant linkages of single markers, as these results were probably caused by allele frequency skews.

We genotyped the *zeel-1 peel-1* and *plg-1* loci using genomic PCR for each of the 200 strains (**Supplementary Table 2**). These presence or absence of genotypes might not reflect the phenotypes for these variants.

For assessment of abamectin sensitivity, *C. elegans* in the L4 stage were grown for 20 h on NGM plates freshly seeded with *E. coli* OP50. Young adult worms were then transferred onto an unseeded plate and allowed to roam for 1 min, then were transferred one per well into a 96-well flat-bottom tissue-culture–treated microtiter plate (Costar) containing 150 µl of M9 buffer with 5 µg/ml abamectin. Worms were monitored at room temperature (~22.5 °C) under a Leica SMZ650 dissecting scope to measure body bends in a 10-s period, either by direct observation or by video recordings. A single body bend was defined as bending on either the dorsal or ventral side relative to the midline.

For assessment of *P. aeruginosa* avoidance, we scored the fraction of worms that crawled off the agar plate during a slow-killing assay. Slow-killing assays were performed as previously described[39]. Briefly, the standard slow-killing assay[40] was performed in the presence of 50 µg/ml 5-fluorodeoxyuridine using the PA14 strain. A minimum of 80 worms per genotype were assayed in at least two independent trials.

Determination of segment sharing. We ran the program GERMLINE[26] on the imputed 40,857-SNP set to define shared segments as intervals of at least 150 markers and 2 cM or Mb in length, with no more than two SNPs between isotypes. Shared segments were collapsed into a single haplotype, and we calculated the haplotype frequencies and homozygosity of each interval in the genome.

Calculation of population genetic statistics. To reduce the effects of sequencing errors on standard population genetic statistics (π, θ_W and Tajima's D), we excluded all singletons and calculated Achaz's Y[41] instead of Tajima's D. Although our error rate was low on a genome-wide scale (less than one false SNP per 10 kb), errors may still have accounted for a large fraction of observed variants in low diversity regions of the genome, substantially biasing the Tajima's D value[42].

We estimated the population recombination rate ($\rho = 4Nr$) for each chromosome using composite likelihood[20] as implemented in LDhat (version 2.1) using values of ρ between 0 and 250 in increments of five. The outcrossing rate C was estimated as

$$C = \frac{\rho}{4Nr_c}$$

where $r_c = 0.5$ is the recombination rate per chromosome per outcross, with the effective population size assumed to be between 10,000 and 50,000 individuals.

Simulation of population genetic parameters. We performed coalescent simulations of entire chromosomes using the program msms[43]. To match the observed recombination patterns, we adjusted the arms of the simulated chromosomes by dividing the distance between each pair of SNPs by five (thereby increasing the effective recombination rate fivefold) and randomly removed SNPs to maintain SNP density. SNPs in the center of the chromosome were randomly removed with probability of 0.9 to match the observed patterns of polymorphism without affecting allele frequencies. The final chromosomes thus contained three regions: two high-diversity regions, high-recombination arms covering 20% of the physical chromosome each and a central region with low recombination and low diversity. The total chromosome length was set at 17 Mb. For all simulations, the population mutation rate (θ) and the recombination

rate (ρ) were uniformly sampled across a broad range of values (θ = U (4,000, 20,000), before the reductions described; ρ = U (50, 250)). Simulated chromosomes with a calculated θ_W (singletons excluded) in the range of the observed data (700–1,150) were accepted, and 10^6 such chromosomes were generated for every set of models.

Simulations with selection consisted of a single population with a single selected site in the chromosome center with a final frequency in the population of 90%. We sampled from logarithmic distributions for both selective coefficients ($\log_{10}$ ($4Ns$) = U (−2, 6)) and population sizes ($\log_{10}$ (N) = U (2,6)). Simulations in which the calculated values of Achaz's Y and the average haplotype homozygosity differed from the observed value by less than 0.05 were used to construct a distribution of possible values of $4Ns$.

Neutral simulations included models of a single population with constant size or a recent period of exponential growth. Models with two ancestral populations were also considered, with individuals sampled from each population in proportion to their relative sizes. Parameters of this model (in addition to θ and ρ) included rates of migration between populations (migration could be asymmetric and change over time) and the relative population sizes.

Coalescent simulations to determine the age of the chromosome V haplotype. We modeled expansion of the largest highly shared segment, chromosome V at 9.6–11.9 Mb, using coalescent simulations of 84 individuals and no recombination. To model exponential growth, we sampled from uniform distributions of θ and the growth rate parameter, α (with N_t=$N_0 e^{-\alpha t}$, where N_t is the population size $4Nt$ generations in the past and N_0 is the present size). Values for θ and α were retained from simulated samples with 66–68 segregating sites (the observed number of sites was 67) and a Tajima's D between −2.66 and −2.68 (with an observed value of −2.67) (data from the entire population suggest that the Tajima's D value is minimally biased in this region). Using the laboratory-derived SNP mutation rate of 9×10^{-9} per bp per generation[44] to estimate the population size, the median population expansion rate was 0.86% per generation (90% credible interval 0.63–1.4%; **Supplementary Fig. 8**). For a 1,000-fold population expansion, we then estimated a median of 807 generations (90% credible interval 636–1,081).

An alternative simulation approach forced all lineages to coalesce at a given time t (measured in $4N$ generations). For these 'star-like' simulations, we randomly sampled from uniform distributions of θ and t, retaining successful samples as described above. Again using the laboratory-derived mutation rate to estimate population size, we estimated the time to the forced coalescence as 846 generations (90% credible interval 630–1,158) (**Supplementary Fig. 8**).

31. Brenner, S. The genetics of *Caenorhabditis elegans*. *Genetics* **77**, 71–94 (1974).
32. Kiontke, K. *et al*. *Caenorhabditis* phylogeny predicts convergence of hermaphroditism and extensive intron loss. *Proc. Natl. Acad. Sci. USA* **101**, 9003–9008 (2004).
33. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
34. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
35. Roberts, A. *et al*. Inferring missing genotypes in large SNP panels using fast nearest-neighbor searches over sliding windows. *Bioinformatics* **23**, i401–i407 (2007).
36. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
37. Patterson, N. *et al*. Methods for high-density admixture mapping of disease genes. *Am. J. Hum. Genet.* **74**, 979–1000 (2004).
38. Kang, H.M. *et al.* Efficient control of population structure in model organism association mapping. *Genetics* **178**, 1709–1723 (2008).
39. Reddy, K.C., Andersen, E.C., Kruglyak, L. & Kim, D.H. A polymorphism in *npr-1* is a behavioral determinant of pathogen susceptibility in *C. elegans*. *Science* **323**, 382–384 (2009).
40. Tan, M.W., Mahajan-Miklos, S. & Ausubel, F.M. Killing of *Caenorhabditis elegans* by *Pseudomonas aeruginosa* used to model mammalian bacterial pathogenesis. *Proc. Natl. Acad. Sci. USA* **96**, 715–720 (1999).
41. Achaz, G. Frequency spectrum neutrality tests: one for all and all for one. *Genetics* **183**, 249–258 (2009).
42. Achaz, G. Testing for neutrality in samples with sequencing errors. *Genetics* **179**, 1409–1424 (2008).
43. Ewing, G. & Hermisson, J. MSMS: a coalescent simulation program including recombination, demographic structure and selection at a single locus. *Bioinformatics* **26**, 2064–2065 (2010).
44. Denver, D.R., Morris, K., Lynch, M. & Thomas, W.K. High mutation rate and predominance of insertions in the *Caenorhabditis elegans* nuclear genome. *Nature* **430**, 679–682 (2004).

doi:10.1038/ng.1050