Methods

Strains

Animals were cultured on modified nematode growth medium (NGM) plates[1] with 1% agar and 0.7% agarose to prevent increased burrowing associated with all wild *C. elegans* isolates. Strain propagation and preparation for DNA isolation were performed on these plates with a lawn of the bacterial strain OP50.

All wild strains used in this study are 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. The laboratories that generously provided the strains are listed in the table. In all other cases, the *Caenorhabditis* Genetics Center (CGC) provided the strains.

Sequence analysis identified strains for which the true identity is suspect. The CGC versions of strains CB4855 and CB4858 were found to be identical by sequence comparison, even though the strains are reportedly from different isolation locations. 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 the 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; they 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 likely mis-frozen from an RC301 stock and was excluded from our analyses. JU813 and ED3054 were found to be *C. briggsae* by sequence[2] and mating tests, and were not included in any analyses.

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

Restriction-site associated DNA (RAD) marker library construction and sequence determination

We isolated genomic DNA by washing off nearly starved animals from five 10 cm modified NGM plates to 15 mL conical tubes and allowing them to settle by gravity for one hour. The animals were washed two additional times and settled by gravity into microfuge tubes, and the pellet was frozen at -80ºC. Genomic DNA was prepared using the DNeasy Blood and Tissue Kit (Qiagen). DNA concentration was determined using the Broad-range Quant-it fluorimetry kit (Invitrogen). Seventeen of the RAD marker libraries were constructed by Floragenex, Inc. Nine additional libraries were constructed using the following protocol adapted from previous work [3]. 1 µg of DNA was restricted using high-fidelity EcoRI (NEB) for two hours at 37ºC and heat inactivated at 65ºC for 20 minutes. EcoRI does not restrict DNA from the OP50 bacterial food source because of the endogenous *E. coli* modification system, allowing us to avoid sequencing any

contaminating bacterial DNA. 5 μL of 100 nM annealed and barcoded P1 adapter (sequences below) was ligated (Invitrogen) to DNA from each strain for 20 minutes at 25ºC and then heat inactivated for 20 minutes at 65ºC. Ligated DNA samples from individual strains were pooled at equimolar ratios and shearing buffer (Illumina) was added. Samples were sheared to an average size of 500 bp using a DNA nebulizer (Invitrogen). After nebulization, samples were purified and concentrated to 30 μL using QIAquick PCR Purification columns (Qiagen). DNA ends were blunted using Quick Blunting kit (NEB) and purified by QIAquick PCR Purification columns. Adenine was tailed onto DNA ends using Klenow exo- (NEB) and purified using QIAquick PCR Purification columns. P2 adapter was ligated at a 10:1 adapter to DNA ratio for 20 minutes at 25ºC, and then samples were separated using 2% TAE agarose gel. DNA from 200 to 400 bp was excised and extracted using QIAquick PCR Purification columns. Libraries were amplified for 18 cycles using Phusion MasterMix (NEB) and then the sequencing library was eluted using Agencourt Ampure XP magnetic beads (Beckman Coulter). Amplified library concentration was determined using High-sensitivity Quant-it DNA fluorimetry (Invitrogen) and diluted to 10 nM for loading on a Genome Analyzer IIx (Illumina).

P1 top: 5'-ACACTCTTTCCCTACACGACGCTCTTCCGATCTxxxxx-3'
P1 bottom: 5'-
/Phos/AATTxxxxxAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT-3'

P2 top: 5'-/Phos/CTCAGGCATCACTCGATTCCTCCGAGAACAA-3'
P2 bottom: 5'-
CAAGCAGAAGACGGCATACGACGGAGGAATCGAGTGATGCCTGAG*T-3'

Forward amplification primer:
5'AATGATACGGCGACCACCG*AGATCTACACTCTTTCCCTACACGACGCTCT-3'
Reverse amplification primer: 5'-CAAGCAGAAGACGGCATACG*A-3'

/Phos/ is an added phosphate for ligation, and * denotes a thiol linker. [x] is the barcode.

Illumina Genome Analyzer IIx protocols were used for sequencing at 101 cycles for all samples.

SNP determination

FASTQ files were parsed using custom Python scripts, and each sequence read was entered into a custom mySQL database. Reads were grouped by strain, checked for the presence of a complete EcoRI cut sequence, then mapped to the WS210 version of the *C. elegans* genome (strain N2) using *bwa*[4]. For each observed read location (including those sites without an EcoRI site in the reference sequence), we calculated the number of strains with sequence reads at the site and the average number of reads per strain present. Reads from locations with sequence from only a single strain or fewer than five reads per strain were excluded, as were locations less than 100 bp from another cut site. Reads from locations that passed these filters were exported from the database to create pileup

files for each strain using *SAMtools*[5]. These pileup files were parsed for SNPs using custom Python scripts. The SNP information was imported into R, where all subsequent analyses were performed.

SNP calls were filtered to remove those SNPs near repetitive sequence as defined by *RepeatMasker* (www.repeatmasker.org), as well as any SNPs that occurred beyond the expected read length (as would occur in mappings that contained deletions). We empirically determined an appropriate FDR by comparison of duplicate libraries independently generated from different biological replicates of the same strain. Sites that differed between replicate libraries were counted as errors when both calls had quality scores above the chosen threshold. Sites where the calls from both libraries met the chosen threshold and differed from the reference SNP were counted as true SNPs. We found that a phred quality score of 120 provided a very low FDR (~0.6%) without sacrificing power to identify true SNPs (Supplementary Fig. 3). This type of quality-based cutoff for the identification of SNPs may result in a biased estimate of SNP number and population frequency, with the most severe effects on singleton SNPs, which could be missed if they happened to fall in a region of low sequence coverage or quality. By contrast, sequencing errors will tend to inflate the number of singleton SNPs. To minimize such biases, we excluded singleton SNPs from the estimates of population genetic parameters (see below). SNP calls for all strains (including singleton SNPs) have been uploaded to dbSNP and will be available under the submitter handle "KRUGLYAK" (http://www.ncbi.nlm.nih.gov/projects/SNP/snp_tableList.cgi).

SNP set construction

We identified a SNP at any locus where each allele was called with a phred quality score of at least 120 in at least one strain, resulting in 41,188 SNPs. Genotypes were called at a quality score of at least 60 in all other strains. This two-tiered strategy provided the best balance between removing too few spurious calls and sacrificing too many real variants. Any genotype call with a quality score below 60 was considered missing data, and any locus with more than 25 missing SNPs out of the 200 strains was removed. Any tri-allelic SNPs were also removed from the dataset. We found that reducing the genotype-calling quality threshold below 60 reduced the number of SNPs in the final dataset by introducing more tri-allelic SNP calls due to their high probability of being errors. To impute missing genotype calls, we used *NPUTE*[6], testing imputation windows between 10 – 120 markers. We then used the optimum window (as measured by imputation of known SNP calls) for each chromosome, resulting in imputation accuracy of greater than 99.5%. Both the imputed and unimputed datasets were condensed from the 200 strains into the 97 isotypes. This reduction caused a small number of polymorphisms segregating within isotypes to become monomorphic, resulting in a set of 40,857 SNPs of which 16,928 are singletons observed in only one the 97 isotypes. This SNP set was used for all other analyses, except for association analysis and detection of population structure.

For *STRUCTURE* and principal component analysis, we constructed a more stringent SNP set with high-quality genotype calls of common variants. This set contains SNPs present in at least one isotype with a quality score of 120 or greater and called at a quality score of 100 for all other isotypes. SNPs that were missing or low quality in more than

five isotypes were removed. Any SNP with a minor allele count of fewer than six of the 97 isotypes was removed. The remaining missing calls were imputed using the program *NPUTE*, as above. The more stringent cutoffs resulted in a smaller set of 6,089 high-quality common variants.

To construct a SNP set for association mapping, we started with the set of 6,089 high-quality variants used for population structure, but we raised the minor allele cutoff to 10 out of 97 isotypes to focus our analysis on SNPs with higher allele frequencies. This increased frequency cutoff reduced the SNP set from 6,089 to 4,690.

Determination of population structure

For *STRUCTURE* and Principal Component Analysis, we pruned SNPs within 25 marker windows with a sliding window step of five markers, pruning pairs with r^2 greater than 0.3 using the --indep-pairwise function in *PLINK*[7]. This reduced SNP set contained 757 SNPs. We also examined principal components produced from pruning at r^2 thresholds from 0.1 to 0.7. The results of *STRUCTURE* and Principal Component Analysis were similar using different levels of pruning or missing data thresholds. We used *EIGENSOFT*[8] to identify principal components and evaluated significance using Tracy-Widom statistics. Running *EIGENSOFT* with the "missingmode: YES" option confirmed that the population structure we observe is not caused by structure in the missing data.

Association mapping

We also show that association studies are feasible in *C. elegans*, and can be carried out by simply phenotyping the 97 isotypes with distinct genome-wide haplotypes described here. The absence of large-scale population substructure eliminates a confounding factor that affects association studies in maize, *Arabidopsis*, and humans[8-10], and the 97 isotypes can be used for genome-wide association studies with only a kinship matrix needed to account for relatedness within the population.

We genotyped the *zeel-1 peel-1* genomic interval using the primer sets below in two PCRs for each of the 200 strains. The absence of the *zeel-1 peel-1* deletion does not mean that the strain is phenotypically incompatible with strains with wild-type *zeel-1 peel-1*, as some strains have loss-of-function mutations in *peel-1*[11].

A CB4856-like strain with *zeel-1 peel-1* deletion has a longer PCR product than an N2-like strain using the following primers:

oECA241 5'-TGGATACGATTCGAGCTTCC-3'
oECA242 5'-CCCCCTAATTTCCAAGTGGT-3'

An N2-like strain without the *zeel-1 peel-1* deletion has a longer PCR product than a CB4856-like strain using the following primers:

oECA247 5’-CTGAAGCATGCCGGATTTAT-3’
oECA248 5’-TCCGTCCAATATTCAATCGAC-3’

We genotyped the presence or absence of the *Cer1* transposon insertion into *plg-1* using the following primer sets in two PCRs for each of the 200 hundred strains:

Presence of *Cer1* insertion (N2-like)
oECA245 5’-TCCACAAAACCTGCTGACTG-3’
oECA246 5’-ATCCACTCGATTTTCGCAAC-3’

Absence of *Cer1* insertion (CB4856-like)
oECA243 5’-CGCATAAAACGTCAGCAGAA-3’
oECA244 5’-ATTCGGAGTAGTCGGGTCCT-3’

For abamectin sensitivity, abamectin was purchased from Sigma (Catalog # PS2068) and 10 mg/ml stock solutions were made in DMSO and kept at 4ºC in 50 µl aliquots. L4 animals were picked onto NGM plates seeded with *E.coli* OP50 the day before the assay for avermectin resistance. After ~20 hours, young adults were transferred onto an unseeded NGM agar plate and allowed to roam for roughly one minute. Next, animals were individually transferred into wells of 96-well clear, flat-bottom tissue culture treated microtiter plates (Costar). Each well contained 150 µl of M9 buffer with abamectin at 5 µg/mL, and the animals were distributed such that there was one animal per well. The assays were carried out at room temperature. Animals were monitored under a Leica SMZ650 dissecting scope, and the frequency of body bends were determined either by direct observation or video recordings at 30 frames per second over a 10 second period per animal. A single body bend was defined as bending on either dorsal or ventral side relative to the midline. Animals that showed zero body bends over a 10 second duration were defined as paralyzed.

For *Pseudomonas aeruginosa* avoidance, the fraction of animals that crawled off the agar plate during the course of the slow-killing assay was scored. Slow-killing assays were performed as published previously[12]. Briefly, the standard slow-killing assay[13] was performed in the presence of 50 µg/ml 5-fluorodeoxyuridine (FUdR) and using the *P. aeruginosa* strain PA14. A minimum of 80 worms for each genotype were assayed in at least two independent trials.

We used *EMMA*[14] for all association analyses along with the linear model framework containing a kinship matrix to describe the degree of relatedness among strains. The kinship matrix was created using *EMMA*. We also carried out mapping using Kruskal-Wallis or Fisher’s Exact tests of association. Importantly, *EMMA* identified the causal locus similarly to or better than these two tests. We ignored significant linkages of single markers, as these results are likely caused by allele frequency skews.

Determination of segment sharing

We ran the program *GERMLINE* on the imputed 40,857 SNP set. Shared segments were defined as intervals having at least 150 markers with a minimum length of two cM or Mb with no more than two SNPs within a pair of isotypes. The results from *GERMLINE* were parsed in R using custom scripts that divided the genome into segments defined by every segment boundary found by *GERMLINE*. For each of these genome segments, we calculated the number of distinct haplotypes (including those unique to a single isotype), the haplotype homozygosity, and the percentage of isotypes with a particular haplotype. Small segments (less than 10 kb) were not plotted. Moderate changes to the *GERMLINE* initial parameters had little effect on the downstream results, as highly shared segments always appeared in the same locations.

Calculation of population genetic statistics

To reduce the effects of sequencing error on standard population genetic statistics (π, Watterson's θ, Tajima's D), we calculated corrected statistics with singletons excluded[15] using the unimputed 40,857 SNP set from the 97 isotypes. In particular, we used Achaz's Y[15] in place of Tajima's D to measure deviation from the neutral site frequency spectrum. Note that, although our error rate is low on a genome-wide scale (less than one false SNP per 10 kb), errors may still account for a very large fraction of the observed variants in the low diversity regions of the genome, substantially biasing Tajima's D[16].

We estimated the population recombination rate ($\rho = 4Nr$) for each chromosome using composite likelihood[17] as implemented in the *LDhat* package (version 2.1: http://www.stats.ox.ac.uk/~mcvean/LDhat). We calculated the likelihood of the data under values of ρ between 0 and 250 in increments of 5. Estimates of the frequency of outcrossing were calculated as $C = \frac{\rho}{4Nr_c}$ where C is the rate of outcrossing and $r_c = 0.5$ is the recombination rate per chromosome per outcross, with effective population size assumed to be between 10,000 and 50,000.

Simulation of population genetic parameters

To identify the parameters most likely to have shaped the patterns of genomic variation in the *C. elegans* population, we performed coalescent simulations of entire chromosomes with and without selection using the program *msms*[18]. In order to match the gross patterns of variation in recombination rate and polymorphism across *C. elegans* chromosomes, we adjusted the arms of the simulated chromosomes by dividing the distance between each pair of SNPs by five (increasing the recombination rate five fold), and by randomly removing SNPs with a probability of 0.8 to maintain SNP density. SNPs in the center of the chromosome were randomly removed with probability 0.9 to match the observed reductions in polymorphism without affecting the site frequency spectrum. The final chromosome then consisted of three regions: two high diversity, high recombination arms covering 20% of the physical chromosome each, and a central region with low recombination and low diversity covering the central 60%. Calculations of population genetic statistics and segment sharing were then performed as described for the observed

data, with the total length of the chromosome set at 17 Mb. For all simulations, the input population mutation rate (θ) and population recombination rate (ρ) were uniformly sampled across a broad range of values [θ = U(4000, 20000), before reductions described; ρ = U(50, 250) (estimated ρ for chromosomes ranged from 90 – 185)]. Simulated chromosomes with a calculated θ_W (singletons excluded) in the range of the observed data (700 – 1150) 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 center of the chromosome with a final frequency in the population of 90%. Simulations were performed using parameters sampled from a log range of selective coefficients [log10(*4Ns*) = U(–2, 6)] and population sizes [log10(*N*) = U(2,6)]. Simulations where the calculated values of Achaz's Y and average haplotype homozygosity differed from the observed value for a given chromosome by less than 0.05 were used to construct a distribution of possible values of *4Ns*. For chromosomes II and III, the distribution did not depend much on *N*, and was nearly uniform for values of *4Ns* < 100.

Simulations without positive selection included simple models of a single population with constant population size or a recent period of exponential growth. Models with two ancestral populations were also considered, with the two populations joined immediately before sampling (so that individuals came from each population in proportion to the relative population sizes). Parameters of this model (in addition to θ and ρ) consisted of the rates of migration between the populations (migration could be asymmetrical) and the relative sizes of the two populations.

Coalescent Simulations to Determine the Age of the Chromosome V Haplotype

We modeled the expansion of the largest highly shared segment, found on chromosome V at 9.6-11.9 Mb, using coalescent simulations of 84 individuals. Because this region showed no evidence of historical recombination, we treated it as a single locus. To model exponential growth, we sampled from uniform distributions of values for θ and the exponential growth rate parameter, α ($N_t = N_0 e^{-\alpha 4 N_0 t}$, where N_t is the population size t generations in the past and N_0 is the population size at the present). Values for θ and α were accepted for any simulations that generated samples with 66-68 segregating sites (observed = 67) and a Tajima's D between -2.66 and -2.68 (observed = -2.67). Because most of the segregating sites on the shared haplotype are singletons, we used Tajima's D instead of Achaz's Y. The data from the entire population suggest that Tajima's D is minimally biased in this region. As θ and α both depend on population size, we are most interested in the ratio α/θ, which is equivalent to the per-generation population growth rate divided by per-generation mutation rate. Assuming the laboratory-derived SNP mutation rate of 9×10^{-9} per base pair per generation[19], the median population expansion rate is estimated to be 0.86% per generation (90% CI: 0.63 – 1.4%). We can use this range to approximate the age of the haplotype by estimating the number of generations it took for the haplotype to increase in frequency *G* fold as log(*G*)/(α/4*N*). For a 1000-fold population expansion, we then estimate a median of 807 generations (90% credible

interval = 636 – 1081). A 10,000-fold expansion raises the median estimate to 1075 generations.

In the second simulation approach we rapidly 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* and obtained a distribution of the successful parameter space that contained a distinct center at θ = 72, *t* = 0.02 (Supplementary Fig. 20b). Again using the laboratory-derived mutation rate to estimate population size, we estimate the time to the forced coalescence as 846 generations (90% CI: 630-1158). We note that the accuracy of generation time estimates depends on the accuracy of the estimate of the neutral mutation rate, which may vary substantially between laboratory and natural environments.

References

1. Brenner, S. The genetics of *Caenorhabditis elegans*. *Genetics* **77**, 71-94 (1974).
2. Kiontke, K. *et al. Caenorhabditis* phylogeny predicts convergence of hermaphroditism and extensive intron loss. *Proc Natl Acad Sci U S A* **101**, 9003-8 (2004).
3. Baird, N.A. *et al.* Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS One* **3**, e3376 (2008).
4. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-60 (2009).
5. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-9 (2009).
6. Roberts, A. *et al.* Inferring missing genotypes in large SNP panels using fast nearest-neighbor searches over sliding windows. *Bioinformatics* **23**, i401-7 (2007).
7. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* **81**, 559-75 (2007).
8. Patterson, N. *et al.* Methods for high-density admixture mapping of disease genes. *Am J Hum Genet* **74**, 979-1000 (2004).
9. Atwell, S. *et al.* Genome-wide association study of 107 phenotypes in *Arabidopsis thaliana* inbred lines. *Nature* **465**, 627-31 (2010).
10. Yu, J. *et al.* A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat Genet* **38**, 203-8 (2006).
11. Seidel, H.S., Rockman, M.V. & Kruglyak, L. Widespread genetic incompatibility in *C. elegans* maintained by balancing selection. *Science* **319**, 589-94 (2008).
12. 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**, 382-4 (2009).
13. Tan, M.W., Mahajan-Miklos, S. & Ausubel, F.M. Killing of *Caenorhabditis elegans* by *Pseudomonas aeruginosa* used to model mammalian bacterial pathogenesis. *Proceedings of the National Academy of Sciences of the United States of America* **96**, 715-20 (1999).

14. Kang, H.M. *et al.* Efficient control of population structure in model organism association mapping. *Genetics* **178**, 1709-23 (2008).
15. Achaz, G. Frequency spectrum neutrality tests: one for all and all for one. *Genetics* **183**, 249-58 (2009).
16. Achaz, G. Testing for neutrality in samples with sequencing errors. *Genetics* **179**, 1409-24 (2008).
17. Hudson, R.R. Two-locus sampling distributions and their application. *Genetics* **159**, 1805-17 (2001).
18. Ewing, G. & Hermisson, J. MSMS: a coalescent simulation program including recombination, demographic structure and selection at a single locus. *Bioinformatics* **26**, 2064-5 (2010).
19. 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**, 679-82 (2004).