Genetic diversity landscapes in outcrossing and selfing Caenorhabditis nematodes

Anastasia A. Teterina¹,², John H. Willis¹, Matt Lukac¹, Richard Jovelin³, Asher D. Cutter³, Patrick C. Phillips¹

¹ Institute of Ecology and Evolution, University of Oregon, Eugene, OR, USA; ² Center of Parasitology, Severtsov Institute of Ecology and Evolution RAS, Moscow, Russia; ³ Department of Ecology and Evolutionary Biology, University of Toronto, Toronto, Ontario, Canada

Abstract

Caenorhabditis nematodes form an excellent model for studying how the mode of reproduction affects genetic diversity, as some species reproduce via outcrossing whereas others can self-fertilize. Currently, chromosome-level patterns of diversity and recombination are only available for self-reproducing Caenorhabditis, making the generality of genomic patterns across the genus unclear given the profound potential influence of reproductive mode. Here we present a whole-genome diversity landscape, coupled with a new genetic map, for the outcrossing nematode C. remanei. We demonstrate that the genomic distribution of recombination in C. remanei, like the model nematode C. elegans, shows high recombination rates on chromosome arms and low rates toward the central regions. Patterns of genetic variation across the genome are also similar between these species, but differ dramatically in scale, being 10-fold greater for C. remanei. Historical reconstructions of variation in effective population size over the past million generations echo this difference in polymorphism. Evolutionary simulations demonstrate how selection, recombination, mutation, and selfing shape variation along the genome, and that multiple drivers can produce patterns similar to those observed in natural populations. Convolutional neural networks demonstrate the potential for classifying distinct evolutionary scenarios in simulated populations. Distinguishing these forces confidently with the empirical data, however, will benefit from larger population genomic samples from multiple populations and consideration of an even more extensive training set of simulations.
Introduction

Population genomics aims to infer the evolutionary forces and historical processes that have shaped genetic variation within species while considering location-dependent effects generated by chromosomal structure and other genome-wide influences. Patterns of variation along the genome are driven by a range of factors such as natural selection, patterns of reproduction, genome functional organization, mutational and recombinational landscapes, as well as spatial and temporal population dynamics and demographic history. Because most of these factors act and constantly interact, inferring evolutionary scenarios using DNA sequence information within natural populations can be challenging. While many population genetic models incorporate several of these factors ([1], [2], [3], [4]), the enormous complexity of the problem means that there is no single analytical model that encompasses all of these interconnected processes. Even as the scale and quality of individual-level genomic data within natural populations continues to mount, a great deal of work remains to fully integrate the ways that genetic processes manifest patterns at the whole-genome level in combination with evolutionary processes operating within and between populations and species.

Given this challenge, one promising and integrative approach is to use current knowledge about the genetics of a given species alongside evolutionary simulations of a variety of evolutionary scenarios ([5], [6], [7], [8], [9], [10]) so as to generate a series of null models for hypothesis testing of empirical data. Given recent progress in molecular genetic methods, it is possible to obtain high-quality genetic information on genome properties by assembling chromosome-level genomes ([11], [12], [13]), analyzing genome-wide variation in the rate of mutation ([14], [15], [16]) and recombination ([17], [18], [19], [20]), and measuring functional-genomic patterns of activity ([21], [22]), and to then match these features with population-level whole-genome sequence data ([23], [24]). An important emerging approach is deep learning to use such massive and heterogeneous data to capture and classify hidden dependencies and patterns, a machine learning technique that is already finding applications in functional genomics ([25], [26], [22]), medicine ([27], [28]), and population genetics ([29], [30]). Taken together, high-quality genomic data, population theory, individualized null hypotheses from evolutionary simulations, and approaches to learning to identify signatures of evolutionary forces promise to be a powerful tool in population genetics to tease apart the genetic and evolutionary forces that govern genetic variation within and between species.

Variation in the mating system provides one crucial species-specific factor that influences traits, ecology, and population genetic parameters. For instance, self-fertilization as an extreme form of inbreeding acts to reduce the effective population size and the effective recombination rate ([31], [32], [33], [34], [35]), thereby leading to a reduction in
heterozygosity, increased linkage disequilibrium, reduced influence of dominance, and increased variability in evolutionary trajectories due to the enhanced influence of drift and a concomitant reduction in efficiency of selection ([36], [37], [38], [39], [40], [41], [42], [43]). Within animals, the transition from outcrossing to selfing is often accompanied by accelerated reproductive incompatibility and isolation, relaxation of sexual selection and sexual conflict, degradation of mating ability, and the generation of outbreeding depression ([44], [45]). In the context of population genomic analysis, it is the influence of self-fertilization on linkage disequilibrium and the way that it expands the genomic footprint of natural selection that is of particular interest. So the contrast between extreme linkage disequilibrium in self-fertilizing species and natural variability in recombination rate across the genome in outcrossing species provides a unique opportunity to critically examine the interaction between population genetic and transmission genetic processes in shaping molecular variation within species.

*Caenorhabditis* nematodes are primarily outcrossing species with males and females ([46]), with the exception of three predominantly self-fertilizing hermaphroditic species: *C. elegans*, *C. briggsae*, and *C. tropicalis*. The self-fertilizing mode of reproduction appears to have evolved independently within each species and to have done so from an outcrossing ancestor fairly recently ([47], [48]). Sex in *Caenorhabditis* species is determined by sex chromosome dosage (X), with females and hermaphrodites having two copies of the X (XX) and males only one (X0) [49]. Because sex is determined by the absence of the X, males can arise spontaneously from nondisjunction of the sex chromosome during meiosis (~0.1% cases for *C. elegans* [50], [51]), allowing rare outcrossing to occur. *C. elegans*, a model species for behavior genetics ([52], [53], [54], [55]), development biology ([56], [57], [58]), and experimental evolution ([59], [51], [60], [61]), has a very low rate of effective outcrossing in natural populations ([62], [63], [64], [65], [66], [67], [68]) which makes it an especially good model for investigating the consequences of reproductive mode ([69], [70] [71], [51], [72]). From a population genomic point of view, the genomic landscapes of diversity in all three selfing *Caenorhabditis* species are similar, with a consistent pattern among all chromosomes of higher genetic diversity on the peripheral “arm” regions and lower diversity on the central regions of chromosomes ([65], [66], [73], [74]). This pattern closely mirrors the chromosome-level pattern of recombination rate, which is high in chromosome arms and low in centers, with the latter occupying about half of the chromosome length ([65], [75], [76], [74]). To date, however, no comprehensive population genomic data are available for outcrossing *Caenorhabditis* species.

How might outcrossing be expected to affect the distribution of diversity across the genome? In this study, we use a previously constructed chromosome-level assembly ([77]) and a newly generated high-density genetic map to examine whole-genome DNA
sequence diversity for *Caenorhabditis remanei*. *C. remanei* is an obligate outcrossing nematode that has a significantly larger effective population size and levels of molecular polymorphism than selfing species ([78], [79], [80], [81], [82]). To provide a uniform basis for comparison across species, we also reanalyzed a local sample of *C. elegans* from Hawaii (from [83]) to compare it with *C. remanei* using the same set of diversity statistics. In addition to inferring demographic histories, patterns of linkage disequilibrium, and genome-wide patterns of divergence, selection, and the spectrum of nucleotide substitutions, we performed evolutionary simulations under different evolutionary, mutational, and recombinational scenarios to compare patterns of diversity in *C. remanei* and *C. elegans* with theoretical expectations. We then employed convolutional neural networks to classify evolutionary scenarios in simulated and empirical data and discuss the caveats and perspectives of this approach. We find that overall levels of diversity are strongly determined by the mode of reproduction and that finer chromosome-level differences in polymorphism are governed by the interaction of selection, mutation, and recombination, indicating that a full understanding of evolution, demography, and genetic transmission are needed to understand whole-genome evolution.

**Results**

**Recombination landscape of *C. remanei***

We reconstructed the first genetic map for an outcrossing species of *Caenorhabditis, C. remanei*, using 341 F2 individuals from crosses of two inbred lines (PX506 and PX553) generated from isolates collected near Toronto, Canada ([77], Fig 1, S1 Table). Of the 1,399,638 polymorphic sites among the parental strains, on average 106,071 markers were covered by bestRAD-sequencing. Full filtration for informative markers yielded 7,512 total sites across the genome, which in turn were used to construct the genetic map. The total length of the genetic map was 288.72 cM. Of this, 40.99 cM is the X chromosome determined via the female parent, and the remainder is the result of sex-averaged maps for each autosome (see details in S1 Table). The X chromosome has lower rates of recombination than the sex-averaged map of autosomes, which is likely driven by differential control of recombination on the sex chromosome (for example [84], [85], reviewed in [86]), and/or unequal recombination rates in males and females ([87], [88], [89], also reviewed in [90]). Our crossing approach precludes the construction of sex-specific maps, although examining any potential differences in recombination between males and females could be of interest to future studies.

For the purposes of the population genomic analysis, we are particularly interested in the shape of the recombination landscape across the whole genome. We found the
chromosomal recombination landscape to be non-uniform in *C. remanei* in a fashion that is superficially similar to that seen in other *Caenorhabditis* species such as *C. elegans*, *C. briggsae*, and *C. tropicalis* with ends of the *C. remanei* chromosomes having elevated recombination rates and for rates of recombination within a given chromosomal domain (arm or center region) being fairly uniform (Fig 1, [65], [74], [75], [76], [91]). We defined the boundaries of the central domains of lower recombination using stepwise regression (Table 1) and refer to the central domain of lower recombination as “Center” and the peripheral regions of higher recombination as “Arms” in the following analyses. Regions of low recombination in *C. elegans* tend to be somewhat larger in relative size than in *C. remanei*. Specifically, the central domains of chromosomes I, II, III, IV, V, and X of *C. elegans* represent, correspondingly, 48%, 48%, 48%, 52%, 51%, and 36% of the total chromosome (estimated from Table 1 in [65]), while in *C. remanei* these represent 42%, 37%, 34%, 33%, 31%, and 33% of chromosomal length.

![Marey map of genome-wide patterns of *C. remanei* recombination.](image)

Rugged marks on the top reflect the density of markers, while the dashed lines show the boundaries of the central domains (see Table 1). The recombination landscape of *C. remanei* resembles that of other *Caenorhabditis* species, *C. elegans* and *C. briggsae*, with core domains ("centers") of low recombination and peripheral domains ("arms") of uniform and high recombination.

**Table 1. Positions of low recombination domains on chromosomes of the *C. remanei* genome obtained from crosses of the PX506 and PX553 strains.**

<table>
<thead>
<tr>
<th>Chromosome</th>
<th>Left arm ends, Kb (95% CI)</th>
<th>Right arm starts, Kb (95% CI)</th>
<th>Recombination rate, cM/Mb (left, central, and right domains)</th>
<th>Chromosome size, bp</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>5,803 (5,753 – 5,853)</td>
<td>13,007 (12,942 – 13,071)</td>
<td>5.197, 0.673, 3.529</td>
<td>17,247,545</td>
</tr>
<tr>
<td>II</td>
<td>6,828 (6,779 – 6,878)</td>
<td>14,292 (14,060 – 14,524)</td>
<td>3.883, 0.808, 2.938</td>
<td>19,935,723</td>
</tr>
<tr>
<td>III</td>
<td>5,849 (5,828 – 5,870)</td>
<td>11,888 (11,863 – 11,914)</td>
<td>4.791, 0.316, 3.601</td>
<td>17,877,849</td>
</tr>
</tbody>
</table>
Genetic diversity of *C. elegans* and *C. remanei*

In order to compare the landscape of genomic diversity in the outcrossing *C. remanei* to those of primarily self-fertilizing species, we sequenced 14 diploid genomes of *C. remanei* individuals collected in a forest area near Toronto, Canada (S1 Table). To our knowledge, this work represents the first comprehensive analysis of chromosome-scale patterns of diversity for outcrossing species from *Caenorhabditis*. Consistent with patterns observed for selfing species of *Caenorhabditis*, nucleotide diversity is as much as 40% higher in the regions of high recombination than in the central domains on all chromosomes (*C. remanei* mean $\pi \pm SD$ in arms, centers, and total: $1.7 \times 10^{-2} \pm 4.9 \times 10^{-3}$, $1.2 \times 10^{-2} \pm 4.2 \times 10^{-3}$, and $1.5 \times 10^{-2} \pm 5.2 \times 10^{-3}$, see Fig 2). Average levels of polymorphism agree qualitatively with previous estimates from this species based on the analysis of individual genes ([92], [81], [82]).

To directly compare equivalent samples of *C. remanei* and *C. elegans*, we reanalyzed 28 wild *C. elegans* isolates collected at a single location in Hawaii (data from [83], S1 Table) and calculated the diversity landscape across the genome using the same analysis pipeline that we applied to *C. remanei*. Diversity for *C. elegans* was assessed for individual genotypes rather than isotypes, as was performed in the original study, so as to properly retain information regarding genotype frequencies within the population. Consistent with previous reports, this analysis shows that *C. elegans*, like *C. remanei*, has higher diversity levels on the arms compared to the centers ([93], [94], [95], [66], [83]), with mean values ($\pi \pm SD$) in arms, centers, and total of $1.8 \times 10^{-3} \pm 2.4 \times 10^{-3}$, $5.4 \times 10^{-4} \pm 1.2 \times 10^{-3}$, and $1.2 \times 10^{-3} \pm 2.1 \times 10^{-3}$.

The patterns of nucleotide diversity in *C. remanei* are qualitatively similar to *C. elegans* in distribution across the genome but quantitatively different from *C. elegans* in scale by being higher by roughly one order of magnitude. This is consistent with previous observations of substantial reduction of diversity in selfing vs. outcrossing species of *Caenorhabditis* ([92], [78], [96], [79], [80], [97], [81], [82]). Highlighting this point, the number of SNVs that we used in the analysis (after filtering, masking of repeats, regions with low mappability, indels, and their flanking regions) was 243,456 variants for the *C. elegans* sample and almost 10x more, 2,365,750, for *C. remanei*. 

<table>
<thead>
<tr>
<th></th>
<th>IV</th>
<th>V</th>
<th>X</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>8,639</td>
<td>17,108</td>
<td>9,040</td>
</tr>
<tr>
<td>(8397 – 8,881)</td>
<td>(17,072 – 17,145)</td>
<td>(8,960 – 9,120)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1.447, 0.520, 3.615</td>
<td>3.697, 0.270, 2.531</td>
<td>3.047, 0.879, 1.339</td>
</tr>
<tr>
<td></td>
<td>25,790,997</td>
<td>22,502,457</td>
<td>21,501,900</td>
</tr>
</tbody>
</table>
When comparing patterns of polymorphism between domains of high and low recombination, we find that, as measured by $\pi$, chromosome arms are significantly more diverse in general than chromosome centers for both species (C. elegans: $d = 0.62$, $Z = 9.28$, $P$-value < 1e-04; C. remanei: $d = 1.08$, $Z = 13.96$, $P$-value < 1e-04), which also holds when looking specifically within exons and introns (C. elegans exons and introns: $d = 0.52$, $Z = 7.94$, $P$-value < 1e-04 and $d = 0.65$, $Z = 9.73$, $P$-value < 1e-04; C. remanei: $d = 1.36$, $Z = 16.16$, $P$-value < 1e-04 and $d = 0.70$, $Z = 8.33$, $P$-value < 1e-04). Within a given domain, we find that introns are much more diverse than exons within C. remanei (Arm: $d = 0.43$, $Z = -5.34$, $P$-value < 1e-04; Center: $d = 1.36$, $Z = -15.86$, $P$-value < 1e-04) but not significantly different in C. elegans (Arm: $d = 0.06$, $Z = 1.03$, $P$-value = 0.8456; Center: $d = 0.04$, $Z = -0.56$, $P$-value = 0.2927). As described below, this difference across functional groups is undoubtedly caused by the fact that the reduced effective recombination rate is C. elegans makes these domains highly susceptible to selective sweeps and background selection, which homogenizes diversity across linked genetic elements ([81]). In addition, hyper-divergent haplotypes, located mostly in the regions of high recombination, contribute to the difference in diversity among domains ([98]). Consistent with this idea, variance in $\pi$ is nearly twice as large in chromosome arms as in central domains within C. elegans but fairly similar across the chromosome in C. remanei. Genomic patterns of other diversity statistics such as $\theta$, Tajima’s D, variance, skew, kurtosis, the number of haplotypes, H1, H12, H2.H1, ZnS, omega (see for details on statistics), and $\beta$ ([99]) are shown in S2 Fig.
Fig 2. Diversity landscapes of *C. elegans* and *C. remanei*.

Dots represent nucleotide diversity per 100-kb window, lines show locally weighted smoothing of these values, and the vertical dashed lines are the boundaries of regions of low recombination central domain for *C. elegans* from [65] and for *C. remanei* from this study. (A) Nucleotide diversity per 100-kb sliding windows, windows with less than 10% of coverage were removed. Quantitatively, outcrossing *C. remanei* has one order of magnitude greater nucleotide diversity than partially selfing *C. elegans*. However, qualitatively, both species have significantly greater diversity in the regions of high recombination. (B) Nucleotide diversity within exons (yellow) and introns (grey) of protein-coding genes summarized per 100-kb windows, windows with less than 5% of coverage were removed. *C. remanei* shows a large and substantial difference in diversity among exons and introns along the genome, unlike *C. elegans*. This likely indicates that all diversity in *C. elegans* is structured by second-order effects like selective sweeps and/or background selection since both functional and more weakly functional parts of the genome have very similar patterns of variation whereas similar functional elements in *C. remanei* show substantially different patterns of variation.

Divergence and the spectrum of substitutions in *C. remanei*
To facilitate examining the landscape of rates of divergence across the *C. remanei* genome, we reconstructed ancestral states for the *C. remanei* genome by using several genomes of *C. remanei* and a reference genome of *C. latens* ([82], [46], [100]). In so doing, we imputed ancestral states for 56% of the PX506 reference genome (60%, 56%, 58%, 42%, 59%, 68% from the lengths of the chromosome I - X), which included 3,553,584 nucleotide substitutions. All chromosomes had comparable fractions of substitutions from the ancestor (5 ± 0.24%). Calculating the Tamura distance [101] of the PX506 reference genome from the reconstituted ancestral genome shows that overall divergence is 1.6 times greater on the arms than in the central domains (Fig 3A). Divergence tends to be positively correlated with recombination ([102]). While the divergence landscape largely resembles the pattern of nucleotide diversity for the Toronto population of *C. remanei* examined here, chromosome regions showing low divergence tend to be wider than those seen for diversity measures in the regions of low recombination. The divergence time between *C. latens* and *C. remanei* allowed the resolution of population genetic processes (fixation) that are still ongoing among lineages within the Toronto population (segregating variation, Fig 3A, 3C and Fig 2), which provides a more clear definition of the boundaries of the domains. The fluctuations in the ratio of nucleotide diversity over divergence on the boundaries of recombination domains are especially noticeable on chromosomes I, II, III, and V (Fig 3B). These diversity/divergence ratio fluxes could result from selective processes and demographic or non-uniformity in the mutation landscape (Figures S3A and S3B, S3D). The divergence is noticeably lower in exons than in introns, 2.6 times less in the central domains and 2.7 in the arms, and significantly lower in domains of lower recombination (Exons: \(d = 1.13, Z = 54.40, P\text{-value} < 1e-04\); Introns: \(d = 1.63, Z = 69.26, P\text{-value} < 1e-04\); Total: \(d = 0.28, Z = 13.55, P\text{-value} < 1e-04\)) indicating strong selection.
Fig 3. Divergence of the *C. remanei* PX506 strain from ancestral states inferred from the *C. latens* genome.

Each point represents the Tamura 1992 distance on 100-kb filtered windows, with at least 30 kb of the window length containing ancestral states, with lines representing the locally weighted smoothing of these values. (A) Divergence of the reference genome (strain PX506) from the reconstructed ancestral genome. Divergence in central domains is 1.6 times lower than in the arms, and several highly divergent regions are located on the arms of chromosomes II, III, IV, and X. (B) Ratio of nucleotide diversity within the Toronto population of *C. remanei* over divergence shown above. The non-uniform ratio across the genome most likely reflects deviations driven by local variation in natural selection and mutation landscape. (C) Divergence of the reference *C. remanei* genome from the ancestral states in exons (yellow) and introns (grey). Only windows with more than 5 kb of ancestral states within exons or introns are shown. Divergence in exons and introns is,
respectively, 1.7 and 1.6 times less in the central domains than in the arms. (D) The ratio of nucleotide diversity in the Toronto population of *C. remanei* over the divergence estimated in exons (yellow) and introns (grey).

We also used the inferred ancestor to estimate the rate of nucleotide substitution in *C. remanei*, finding that all types of substitutions have a similar pattern of fixation along the chromosomes (S3B Fig). Transition over transversion bias in this comparison is 1.16:1, which is smaller than the 1.5:1 ratio observed within the *C. remanei* population sample from Toronto. Transition substitutions from C→T and G→A were the most common, consistent with previous observations of the *C. elegans* mutation patterns ([103], [104], [105], [106], [98]). The genomic landscape of recombination also has an important effect on the nature of the substitutions, with more C→T and G→A substitutions in the central domain than in the arms, and more C→G and G→C in the arms (S3A Fig, S2 Table). In the *C. remanei* population, the distribution of segregating substitutions along the genome follows comparable patterns to the inferred ancestral patterns (Figures S3B and S3C).

Genome-scale variation in the nucleotide substitution pattern within *C. remanei* could be a consequence of the interaction of recombination, selection, and demographic processes but could also be generated by variation in the mutation process itself. Experiments with mutation accumulation lines in *C. elegans* have shown a 1.2-1.6 times higher rate of base substitutions on the arms relative to the centers of chromosomes ([103]), probably due to mutagenic effects of recombination through double-strand breaks ([107], [108], [109]). Other differences between these genomic regions also might contribute, such as chromatin organization, gene density and their expression, and transposon activity. Accurately inferring the mutation landscape of *C. remanei* will require further research, including the use of mutation accumulation lines and/or extensive sequencing of parent-offspring trios.

**Population structure, outcrossing, inbreeding, and effective recombination**

Sampling worms from a single narrow geographic location helps to describe fine-scale population organization. Most of the worms in the *C. remanei* sample formed one cluster of related individuals, with a few more genetically distant worms (S4 Fig). This cluster was most likely formed from individuals from a single-family lineage displaying intensive inbreeding. This view is supported by the relatively high *F*<sub>Is</sub> values along the *C. remanei* genome (0.38 ± 0.15, S2M Fig).

The inbreeding coefficient was even higher in the *C. elegans* population (0.86 ± 0.22), as expected for selfing or partially selfing species. This *F*<sub>Is</sub> value corresponds to a 92.5%
selfing rate ([110]; \( s = 2F_{is}/(1 + F_{is}) \)) under the assumption of the equilibrium), which falls within the range of values estimated for other *C. elegans* samples (Table 3). The effective selfing rate estimated from interchromosomal linkage disequilibrium (LD) is three orders of magnitude lower, 99.998% ±0.005% (Table 3). This *C. elegans* sample consisted primarily of individuals derived from several distinct genetic lineages, which have been combined into “isotypes” in [66] and [83] and the CeNDR database (https://www.elegansvariation.org, S4 Fig).

**Table 3. Outcrossing rate of *C. elegans* samples reported in different studies.**

The Method column specifies the approach used to estimate the outcrossing rate.

<table>
<thead>
<tr>
<th>Outcrossing rate</th>
<th>Method</th>
<th>Study</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1.3 \times 10^{-2})</td>
<td>Heterozygosity</td>
<td>[94]</td>
</tr>
<tr>
<td>(2 \times 10^{-1})</td>
<td>Heterozygosity</td>
<td>[111]</td>
</tr>
<tr>
<td>(1.7 \times 10^{-2})</td>
<td>Heterozygosity</td>
<td>[62]</td>
</tr>
<tr>
<td>(7.5 \times 10^{-2})</td>
<td>Heterozygosity</td>
<td>This study</td>
</tr>
<tr>
<td>(1.3 \times 10^{-4}, 5 \times 10^{-5})</td>
<td>Linkage disequilibrium</td>
<td>[94]</td>
</tr>
<tr>
<td>(1.6 \times 10^{-5}) to (2.2 \times 10^{-3})</td>
<td>Linkage disequilibrium</td>
<td>[63]</td>
</tr>
<tr>
<td>(&lt; 1.1 \times 10^{-4})</td>
<td>Linkage disequilibrium</td>
<td>[73]</td>
</tr>
<tr>
<td>(2.4 \times 10^{-5})</td>
<td>Linkage disequilibrium</td>
<td>This study</td>
</tr>
</tbody>
</table>

Patterns of genome-wide linkage disequilibrium were drastically different in the *C. elegans* and *C. remanei* samples. *C. elegans* had very large blocks of LD both within and across chromosomes, and LD decays slowly along the entire length of the chromosome (Fig 4 and S5 Fig), which agrees with theory ([41]) and previously reported results ([73]). The degree of LD also varies significantly across chromosomes in *C. elegans* (\( d = 0.58, Z = -151.1, P\)-value < 1e-04), which is also consistent with previous observations ([62], [63]). In contrast, LD within the *C. remanei* population decayed rapidly, within a few hundred base pairs on autosomes and a bit more gradually on the X chromosome (Fig 4), consistent with previous sub-genomic observations for this species ([80]). We observe a background inter- and intrachromosomal LD of 0.1, which is more than a null expectation (an inverse of the sample size, 0.04), that is probably related to the high inbreeding or another demographic process in the collected sample (S5 Fig). Consistent with these observations, the inferred genome-wide effective recombination rate in *C. elegans* was on average 8.4 times lower than in *C. remanei* (\(3.8 \times 10^{-4} \pm 2.7 \times 10^{-4}\) versus \(3.2 \times 10^{-3} \pm 7.7 \times 10^{-3}\); S6 Fig). However, the effective recombination in *C. elegans* did not follow the recombination domain structure. Here, the difference in actual and effective recombination rates among species is clearly driven by the increase in linkage disequilibrium caused by self-fertilization.
Fig 4. The decay of linkage disequilibrium (LD) along the chromosomes of *C. elegans* and *C. remanei* populations.

The median values were estimated from LD between all biallelic polymorphic sites in *C. elegans* and every tenth site for *C. remanei* within either 100-kb windows (overall plots) or 100-bp windows (*C. remanei* inset plot). Each line represents a chromosome.

Demographic history of *C. remanei* and *C. elegans* samples

Inference of demographic dynamics within both species reveal dynamic changes in population size over time for each (Fig 5). The first striking difference is that the estimated effective population size of our *C. remanei* sample is approximately two orders of magnitude higher than that of *C. elegans*, spanning thousands of generations (Table 4). The size of the *C. elegans* sample from Hawaii has changed dramatically in recent generations, which likely can be attributed to the intricate metapopulation structure ([94], [111]). From a historical point of view, the *C. elegans* sample shows a pattern of a precipitous decline in the effective population size, as previously noted ([73]). In contrast, the Toronto population of *C. remanei* studied here has maintained a consistent, relatively large effective population size but also displays notable fluctuations in size over time both in the past and recently (Fig 5).

Table 4. Effective population size ($N_e$) for *C. elegans* and *C. remanei* samples from different studies.

The Method column indicates the approach used to estimate $N_e$. For our study, the 95% confidence intervals for the mean $N_e$ are shown in parentheses.

<table>
<thead>
<tr>
<th>Species</th>
<th>$N_e$</th>
<th>Method</th>
<th>Study</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>C. elegans</em></td>
<td>$2 \times 10^2 - 9.5 \times 10^3$</td>
<td>Allele frequencies</td>
<td>[94]</td>
</tr>
<tr>
<td></td>
<td>$10^2 - 10^4$</td>
<td>Allele frequencies</td>
<td>[93]</td>
</tr>
<tr>
<td>Allele frequencies</td>
<td>Coalescent, low accuracy</td>
<td>Allele frequencies</td>
<td>Allele frequencies</td>
</tr>
<tr>
<td>-------------------</td>
<td>-------------------------</td>
<td>-------------------</td>
<td>-------------------</td>
</tr>
<tr>
<td>$8 \times 10^4$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$10^4 - 10^4$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$6.5 \times 10^3$,</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$(4.2 \times 10^3, 8.3 \times 10^3)$</td>
<td>This study</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$3.4 \times 10^4$,</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$(3.1 \times 10^4, 3.7 \times 10^4)$</td>
<td>This study</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>C. remanei</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$1.6 \times 10^5$</td>
<td>Allele frequencies</td>
<td>$1.6 \times 10^5$</td>
<td>Allele frequencies</td>
</tr>
<tr>
<td>$2.5 \times 10^5$,</td>
<td>Allele frequencies</td>
<td>$2.5 \times 10^5$,</td>
<td>Allele frequencies</td>
</tr>
<tr>
<td>$(2.2 \times 10^5, 2.8 \times 10^5)$</td>
<td>This study</td>
<td>$(2.2 \times 10^5, 2.8 \times 10^5)$</td>
<td>This study</td>
</tr>
<tr>
<td>$1.4 \times 10^6$,</td>
<td>Allele frequencies</td>
<td>$1.4 \times 10^6$,</td>
<td>Allele frequencies</td>
</tr>
<tr>
<td>$(1.4 \times 10^6, 1.5 \times 10^6)$</td>
<td>This study</td>
<td>$(1.4 \times 10^6, 1.5 \times 10^6)$</td>
<td>This study</td>
</tr>
</tbody>
</table>

**Fig 5. Reconstructed demographic history of C. elegans and C. remanei populations.**

(A) Recent demographic history inferred from linkage disequilibrium data ([112]). The lines represent the geometric mean of population size from 100 replicates based on autosomes. For *C. elegans*, generation time was doubled, and population size was divided by two. Both species had significant population declines in recent generations.

(B) Ancient demographic history based on SFS and LD information ([113]). Calculations are based on 100 bootstrapped replicates using eight individuals from each species, with each line representing one replicate. The grey shadow shows the region of recent demographic history, where estimations are less reliable. In this analysis, we used one generation per year, and scaling of the mutation rate ($\times 0.5$) and coalescent time ($\times 2$) for *C. elegans*. Demography derived from individual chromosomes is depicted in S7 Fig. Population sizes of both species fluctuate significantly over time, which is likely to influence estimates of the coalescent population size. Consistent with this, coalescent population size estimates tend to be one order of magnitude less than the population sizes calculated from polymorphism data (Table 4). The long-term population size of *C. remanei* sample is around two orders of magnitude larger than those of *C. elegans*. 
To look for site-specific changes in diversity that might be indicative of the action of natural selection, we also inferred population history for the *C. remanei* population using reconstructed ancestral states and the Relate framework ([114]). Overall, the pattern of demographic history using this approach was very similar to that reported above, as well as being concordant across all six chromosomes (see S7 Fig and S8A Fig). However, after adjusting for multiple comparisons, these estimates did not identify any genomic region that had been subjected to substantial positive selection in the past (S8C Fig). Nevertheless, estimated p-values deviated from the expectation of a uniform distribution in a fashion that was more pronounced on the arms than on the centers, which could potentially serve as a signal of non-neutral processes (S8D Fig). This also is consistent with significant and large differences across diversity and divergence in exons vs. introns in *C. remanei* due to selection discussed above (Fig 2B and Fig 3C). We attempted a similar analysis for *C. elegans* but could not reconstruct a sufficient number of ancestral sites from several strains of *C. elegans* and *C. inopinata* (data not shown) to allow the analysis to proceed.

**Evolutionary simulations**

The empirical data shows two major features. First, that total polymorphism is lower in the partial-selfing *C. elegans* relative to the outcrossing *C. remanei*, and second, that the genomic landscape of genetic polymorphism is structured and appears to be strongly correlated with domains of high and low recombination. Because many interacting factors can potentially influence these observations, we conducted individual-based simulations to better understand the separate and combined effects of positive selection, background selection, recombination, variation in mutation rates, variation in demography, and variation in rates of partial selfing on population genomic signatures within these species. We looked at three separate scenarios on a wide array of population genetic statistics using evolutionary simulations in SLIM ([9]): the interaction of selfing rate, selection, mutation landscape, and recombination domains; the decay of the ancestral diversity; the effects of fluctuations in population size.
Fig 6. Nucleotide diversity landscapes in simulated populations.
Lines represent nucleotide diversity per 40-kb window, the vertical dashed lines are the boundaries of regions of low recombination central domain. Columns show the outcrossing rate, where “outcrossing” means completely outcrossing populations and, in other columns, % specify the percentage of selfing in population; “outcrossing inbreeding” corresponds to scenarios with outcrossing populations that underwent the bottleneck at the very end of simulations (see Methods). Rows represent domain-specific differences in mutation rate, with 1-1-1 as the uniform mutation landscape, 1.5-1-1.5 as 50% more mutations in domains of high recombination, and 2-1-2 meaning two times more mutations in domains of high recombination. Colors show the selection regime (see Methods). (A) Mean nucleotide diversity per 40-kb sliding non-overlapping windows. The mean values are lower in regions of low recombination in scenarios with selection and a non-uniform mutation landscape. Higher selfing rate and selection pressure reduce the coalescent time and, consequently, nucleotide diversity. (B) The standard deviation of nucleotide diversity in scenarios with uniform mutation landscapes. The variance in diversity gets lower with selection and reduction of effective recombination in non-neutral conditions.
scenarios and becomes higher with increased selfing in the case of neutrality ([115], [116]).

As predicted by theory ([31], [32], [33], [34], [117], [43]), selfing dramatically reduces the diversity in the population, especially when combined with either positive or negative selection (Fig 6A, far right panels). Because recombination has little impact on within-lineage diversity under self-fertilization, any form of selection tends to eliminate the variation at linked sites, often at the scale of the whole genome. As a corollary of this, variation in recombination rate across the genome exerts little influence on the genomic landscape of polymorphism when the selfing rate is high because the genetically effective recombination rate becomes very low. Similarly, the expected variance in evolutionary outcomes is also very low when selfing and selection combine because selection consistently eliminates variation irrespective of when and where new mutations arise within the genome (Fig 6B).

As even minor amounts of outcrossing enter the population ([118], [119], [120], [121]), however, the situation changes dramatically. In the neutral case, the genomic landscape of polymorphism remains flat regardless of outcrossing rate, as predicted (Fig 6A, [33], [35], [41]). But when selection is introduced, regardless of whether it is positive, negative, or balancing, then regions of high recombination maintain substantially more variation than regions of low recombination (Fig 6A, left most panels). These simulations, in particular, do a very good job of recapitulating the empirical patterns seen in both C. elegans and C. remanei. Recent inbreeding, as we see in our C. remanei samples, does not strongly disrupt this pattern, although the influence of recombination on the genomic distribution in polymorphism is reduced under balancing selection in face of inbreeding, as might be expected. Importantly, however, balancing selection does not generally lead to qualitatively different genomic patterns of polymorphism, nor does it change expectations of the influence of selfing on regions of high and low recombination (see also Figures S10, S11). The variance in outcomes among simulations also displays a recombination-dependent pattern (Fig 6B). Specifically, neutral scenarios have greater variance in regions of lower recombination ([116]). Whereas the variance of π in those domains in non-neutral simulations is lower than in neutral ones and extremely low in scenarios with positive selection.

Although variation in recombination does an excellent job of capturing the genomic differences in polymorphism observed in our empirical examples, it is also possible that these patterns could be caused by domain-specific variation in mutation rate. Indeed, there is evidence that mutation rates are different among recombination domains in C. elegans ([106], [103]). If mutation is in fact elevated on the chromosome arms, then
polymorphism also increases on those arms, with the difference between arms and centers increasing with the disparity in mutation rate (see Fig 6). As long as there is natural selection and a sufficient level of outcrossing, then changing the mutational landscape does not qualitatively alter expectations of the pattern of genomic variation with respect to the influence of domains of high and low recombination. The situation is more complex under neutrality and/or with a high degree of selfing, in which case the disparity in mutation rate can mimic the pattern of polymorphism expected under the combination of selection and recombination. Thus, while the pattern of variation in *C. elegans* appears to be consistent with some level of outcrossing combined with selection, it is also consistent with selfing and variation in mutation rate correlated with recombination (Fig 6). As outlined below, distinguishing these cases, therefore, depends on the combined effects of all of these factors on additional haplotype and site-based statistics, such as theta, the number of haplotypes, variance, kurtosis, H12, H1.H1, H1, \( \beta \)-statistics, \( F_{is} \), omega, that are affected by the mutation landscape and other statistics, such as Tajima’s D and LD based statistics (ZnS), that tend not to be (Figures S10, S11, S12).

Most of the above-mentioned diversity statistics reflect recent evolutionary processes in a population. With simulations, we can explore more distant evolutionary past. One of the statistics that reflects the ancestral relatedness of individuals is the time to the most recent common ancestor (TMRCA). For outcrossing populations, TMRCA is smaller in the regions of lower recombination and reduces in scenarios with selection, with a greater reduction in regimes with beneficial and deleterious mutations. The TMRCA landscape inferred for the Toronto sample of *C. remanei* (S8B Fig) has a distinct shape which may be connected to the population structure and history as well as potential errors in the inference of old tree branches. Selfing, as expected from the theory ([33], [35]), drastically decreases TMRCA due to a reduction of the population size and effective recombination. The variation in the mutation landscape, naturally, has no effect on the TMRCA landscape.

To go further deep in time, we can examine the patterns of divergence from the ancestral genomic state in simulated populations by sampling a random individual at the end of a simulation and comparing its genome to the ancestral "blank" genome (S10O Fig). The divergence is higher in the regions of higher recombination in outcrossing simulated populations and populations with partial selfing, and elevated mutation rate naturally adds on to the divergence, but the reproduction mode has no effect on the divergence pattern. Populations with non-neutral selection regimes show non-uniform ratios of diversity/divergence along the genome. We implemented only simple scenarios with one population to roughly cover the subject of the relationship of diversity and divergence using a single genome comparison and diversity in one population analogous to what we
did with empirical data (Fig 3), to investigate nuances of the relationship of diversity and divergence simulations with several populations are required.

**Fig 7. Decay of ancestral polymorphism in simulated populations that transitioned from outcrossing to selfing.**

The columns indicate populations that switched to 98% selfing (left column) and 100% selfing (right column) after obligate outcrossing. Lines represent the mean nucleotide diversity values estimated from 50 replicates in domains of high recombination, with the shaded areas representing the standard deviation among these replicates. Generations on the ordinate axis show the number of generations after the burn-in. Colors show the selection regime. Each row represents the mutation landscape (see Methods), and the dashed lines represent the nucleotide diversity in the populations without outcrossing ancestors. The average diversity in all scenarios with selection decline rapidly (within ~N generations) towards levels of diversity in selfing or partially selfing populations. Neutral scenarios in obligate selfing populations have greater variance, which increases over time because these populations are, basically, composed of independent lineages.

Another formal possibility for the genomic pattern observed within *C. elegans* is that, while we might expect this species to exhibit a pattern consistent with selfing, there might be residual, *C. remanei*-like unresolved ancestral variation on the arms due to its transition from outcrossing to selfing (see [98]). To examine this possibility, we used evolutionary
simulations to explore the decay rate of ancestral polymorphism from an outcrossing ancestor to a population experiencing either 98% or 100% selfing. For complete selfing and purely neutral variation, the ancestral pattern of variation can indeed persist even after $6N_e$ generations (Fig 7). However, the addition of any form of natural selection and allowing for a small fraction of outcrossing individuals leads to the rapid decay of ancestral diversity, within $1N_e$ generations for 100% selfing and $2N_e$ for 98% selfing. These results are consistent with the expected average coalescent time for the scenario ($\sim 1/N$, more specifically $(1+F)/2N$, where $F$ is the selfing rate; [31], [32], [33], [35], [41], [122]). Here, the variance in the nucleotide diversity is considerably higher in the populations that transition to obligate selfing, which also agrees with previous studies ([123], [35]).

The above simulations assume populations of constant size, but we know that variable sizes are the reality for most populations, especially those that might experience more of a “boom and bust” life cycle, such as that experienced by these species ([124]). We, therefore, also conducted a third type of simulation that examined two types of recent change in the population size under neutral scenarios: 1) change from 15K to 5K every five generations for 100 generations after the burn-in; 2) exponential growth by 3% every generation for 100 generations. We then evaluated the diversity statistics and compared them for each simulation after 100 generations of changes in the population size and at the burn-in. S13 Fig illustrates the fold change in a broad range of population genetic statistics in neutral simulations, with large effects in different directions in many of the statistics used in the study, but especially notable in Tajima’s D, haplotype-based skew, and kurtosis. Importantly, these statistics are sensitive to perturbations in population sizes, which occur naturally in wild worm populations. Likely, the impact of recent demographic shifts on diversity might be less pronounced in the regions of low recombination with selection as the variance in those scenarios is lower (e.g., Fig 6B). Thus, understanding population structure and local dynamics in a population is particularly valuable for interpreting observed patterns of diversity.
Fig 8. Classification of the empirical landscapes of *C. elegans* and *C. remanei* by evolutionary scenarios.

The predictions show the probability (softmax) of each bootstrapped replicate of a chromosome to belong to a specific class of evolutionary scenarios. Only classes with more than 5 in total are shown. Colors represent the bootstrapped replicates of chromosomes; for each nematode, we used only 4 autosomes. Columns correspond to the selfing rate, where “outcrossing inbreeding is outcrossing populations that underwent a bottleneck, and the percentage of selfing defines the selfing rate in simulated populations. The classes represent mutation landscape (“1” – uniform; “1.15” – 15% more mutations on the arms; “1.5” – 50% more mutations on the arms; “2” – 2 times more mutations on the arms. “N”, “ND”, “NDB” mean neutral, neutral, and deleterious, and neutral, deleterious, and beneficial selection regimes. All of the predictions indicate some amount of outcrossing in *C. elegans*. *C. remanei* was not classified as an outcrossing species, which it undoubtedly is. Misclassification might be caused by the complex demographies of nematode populations and/or incompleteness of genomic data, especially on the arms.

Classifications

Because multiple genetic and evolutionary processes can lead to similar outcomes in terms of genomic diversity (e.g., mutation vs. recombination), we sought a comprehensive approach toward distinguishing major factors structuring genetic variation. We, therefore, trained a convolutional neural network on nine diversity statistics (π, θ, Tajima’s D, variance, skew, kurtosis, ω, and β -statistics) with the intent to classify population characteristics such as selfing rate, mutation landscape, and selection, and then apply this network to our empirical samples for *C. elegans* and *C. remanei*. The network was
trained on the first set of simulations described above, with statistics estimated on 40-kb genomic sliding windows and normalized for each simulation. Unfortunately, haplotype-based statistics had to be excluded because they could not be scaled using fixed window size approaches given the drastically different estimated effective population sizes of C. elegans and C. remanei. Additionally, we excluded simulations with balancing selection as they were mostly classified as outcomes including deleterious and beneficial mutations.

Overall, the network did remarkably well classifying simulation results under most parameter combinations (S14 Fig). Matthews correlation coefficients that show the accuracy of predictions of unbalanced classes for joint-feature predictions was 51.8%, with specific predictions for mutation landscape being 67.7%, selfing rate being 85.2%, and selection regime being 71.7%. The confusion matrix is in S15 Fig. The largest point of error for the network was generated by subtle differences in mutation rate across chromosomal regions, with the network confusing the uniform mutation rate with the 15% increase in mutation rate on the arms (S15 Fig). Many of the population genetic statistics are quite similar for these scenarios (S10 Fig). Other cases that confused the network are distinguishing between simulations with neutral mutations and a combination of neutral and deleterious mutations. This is undoubtedly because even in the deleterious mutation case, many of the mutations are drawn from a distribution in which about a third of the selection coefficients have $N_s < 1$ (S9 Fig). Also, many deleterious mutations tend to have short persistence times, so those present in a given population tend to be recessive and effectively neutral when rare.

Then we classified the normalized diversity patterns from the empirical data from C. elegans and C. remanei populations using the network trained on the simulations (Fig 8). For C. elegans, predictions tended to differ per chromosome. In each case, the selfing rate was predicted to be less than 100%, and some amount of effective recombination was detected, which is consistent with our estimation of outcrossing rate directly from linkage disequilibrium. For whatever reason, Chromosome I from this population of C. elegans has a qualitatively different pattern of variation that is more consistent with frequent historical outcrossing and more recent self-reproduction. One of the many possibilities is that it may be connected to the zeel-1/peel-1 system, located on Chromosome I, which is responsible for gametic incompatibility and known to be under balancing selection ([125]). For C. remanei, all chromosomes were unexpectedly classified as adhering to one of the selfing scenarios. This is surprising because the network performed very well in distinguishing among outcrossing rates in the simulations.

To explore these differences a bit more fully, we analyzed the distribution of the simulation and empirical sample statistics using principal component analysis (S16 Fig). In keeping
with the above, positions of the *C. elegans* chromosomes in the principal component space tended to be dispersed but within the range of simulated outcomes. In contrast, *C. remanei* tended to be located close together, indicating the similarity of population genetic processes but were at the edge of the scenarios captured by the simulations (S17 Fig). This would seem to indicate other demographic processes not included here have influenced the subtle signals being captured by the population genetic statistics. This conclusion is reinforced by the fact that saliency maps for the network (S18 Fig) show that even small perturbations in theta, kurtosis, and skew in the central domain are enough to change the predictions for outcrossing simulations. That means that the neural network is sensitive to small differences in the ratio of diversity among domains of recombination.

The evolutionary simulations examining recent fluctuations of population size discussed above show that these fluctuations can in principle generate large changes in most of the diverse statistics used in the network classification. It is therefore likely that the main confounder for the *C. remanei* network analysis is demography (see S13 Fig and Fig 5). Unfortunately, there is an effectively infinite number of possible demographic scenarios involving population size fluctuations, and we did have enough computational resources for full exploration of these models at full population size, as such simulations are extremely memory and time-consuming. Refinement of the network model thus awaits additional empirical information and/or an expanded analytical framework.

Finally, at the whole genome level, choices made during bioinformatic processing can also potentially subtly affect diversity statistics in a way that the network might not see. For instance, when filtering the sequence data, even small changes in filtering parameters, like changing the threshold for invariant sites can shift most of the diversity statistic estimations for *C. remanei* (S19 Fig).

**Discussion**

The transition from traditional population genetics to molecular population genetics shifted the general analytical framework from alleles at a locus to the nucleotide sequence at a particular site in the genome. The progressive expansion of this framework requires scaling these approaches in the context of broader genome-wide factors such as linkage, recombination, and localized variation in mutation rates, as well as how the impacts of these processes are amplified by population history and structure, and species biology. Attempting to understand the separate and combined impacts of these factors requires comprehensive information about molecular diversity across the genome and a theoretical context in which different alternatives can be rigorously tested. *Caenorhabditis*
nematodes provide a natural experiment in which phylogenetically close species have drastically different lifestyles, demography, and genetic processes. With this in mind, we assessed the genomic landscape of various population diversity statistics of populations of *C. remanei* and *C. elegans*. We find that the level of diversity is dramatically different across species, with partially selfing *C. elegans* having order of magnitude lower diversity than the outcrossing *C. remanei*, yet that the pattern of genetic diversity is strikingly similar across both species’ genomes, being positively correlated with large-scale transitions in recombination rate between chromosome centers and arms.

**Mating systems and global patterns of diversity**

Focusing first on global differences among species, here we used local samples of individuals to facilitate analysis of population history and structure and to allow comparison of features of diversity distributions among outcrossing and selfing species with similar genome organizations. Previous analysis of diversity of *C. elegans* at a global scale ([83], [98]) has demonstrated that even with an overall low genetic diversity expected for a nearly selfing species, a large number of hyper-divergent haplotypes—covering one-fifth of the genome and located mostly on the chromosome arms—have been maintained in that species, possibly because they are involved in adaptation to specific environmental niches. These regions also could result from allelic sequence divergence, analogous to the Meselson effect ([126]) demonstrated for asexual species, expected from coalescent theory when the fraction of outcrossing is considerably lower than the reciprocal of the population size ([118], [119], [122], [127]). Overall, our estimations of the outcrossing rate and population size for the *C. elegans* sample overlap with previously reported values (Tables 3 and 4). It remains an open question if such hyper-divergent regions are present in populations of outcrossing *Caenorhabditis* species, because short-read data, such as that used here for *C. remanei* tends to be difficult to align in hyperdivergent sections of regions of high recombination rate. Long-read data will be required to close these gaps in the coverage.

Looking broadly at haplotype structure across the genome, the rate of decay in linkage disequilibrium (LD) is drastically different in *C. remanei* and *C. elegans*. As predicted by theory ([128], [41]), the outcrossing *C. remanei* shows rapid LD decay, within hundreds of base pairs, and low levels of inter and intrachromosomal correlations. These observations and our estimations of genome-wide effective recombination rate are consistent with previous studies ([80]). However, slightly elevated background LD could indicate a recent inbreeding, consistent with the inferred demographic history (Fig 5) and high *F_{IS}* values (S2M Fig). The “Scottish tartan” pattern of genomic linkage disequilibrium in *C. elegans*, in which associations across chromosomes can often be nearly as strong
as associations within chromosomes (Fig. S5), is particularly striking and is consistent with the previous observations of high inter- and intrachromosomal linkage disequilibrium in C. elegans ([80]). Because effective recombination rate is a major driver of diversity and linkage along the genome, nematodes provide an especially useful model for studying the evolutionary effects of this factor.

Genomic landscapes of diversity and factors that affect them in C. elegans and C. remanei populations

Moving from a global genome perspective to patterns of diversity observed along specific chromosomes requires teasing apart a large number of potentially influential factors (including population history, structure, and demography, natural selection, species biology and development, and reproduction mode, location-specific variation in recombination and mutation, and additional genomic properties such as genome activity and positions of various genetic elements like transposons, genes, regulatory elements, etc.). Using the empirical data from populations of C. elegans and C. remanei and evolutionary simulations derived here, we step through below how some of these factors contribute to the landscape of diversity.

Recombination landscape and genomic organization

We constructed the first comprehensive genetic map of C. remanei, the first for an outcrossing species in this genus (Fig 2), to guide the understanding of genomic patterns of diversity. The genetic map has a similar structure to other maps constructed for Caenorhabditis species ([65], [75], [74]) with large central parts of chromosomes of suppressed recombination ([65], [75], [91]). The recombinational landscape deduced from the population data was also consistent with this pattern. Notably, the X chromosome exhibits the same recombination pattern as autosomes in C. remanei, as in other Caenorhabditis species ([65], [75], [91]), but with a noticeably shorter genetic map. More research is needed to determine if this is due to distinct recombination regulation on the sex chromosome ([84], [129], [130], [90]) or to averaging of recombination rates between sexes, which might have different recombination regulation, as has been shown for C. elegans ([131], [90]). C. elegans and C. remanei, and some other Caenorhabditis species, share other similarities in chromosome organization: central parts of chromosomes display lower recombination and have higher gene density, lower repetitive content, lower GC-content, higher gene expression, and a higher level of inter-chromosomal interactions than the peripheral parts of chromosomes ([132], [133], [77]). As recombination across the genome is one of the critical factors that might affect the

CC-BY-NC 4.0 International license available under a which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made
available under aCC-BY-NC 4.0 International license.
shape of diversity landscape, and the genetic map is necessary to model and interpret
diversity along the genome, we used the domain-like structure of *Caenorhabditis*
chromosomes as the basis for our simulations to allow comparison of diversity
landscapes between domains in empirical and simulated populations.

**Diversity landscapes and linkage-disequilibrium**

Self-reproduction leads to increased homozygosity and linkage disequilibrium, which is
known from theory ([31], [32], [33], [35], [134], [135], [136], [122]) and which has been
frequently shown empirically (reviewed in [137]). For *Caenorhabditis* nematodes,
genomic landscapes of diversity have already been derived for three selfing species ([66],
[73], [74]). In order to compare the diversity patterns of nematodes with different
reproduction modes, we obtained the first genome-wide diversity landscape for an
outcrossing *Caenorhabditis* species, *C. remanei*, and contrasted its patterns with the
selfing *C. elegans*. *C. remanei* has one order of magnitude higher nucleotide diversity
than *C. elegans*, due to larger population size, higher effective recombination, and
outcrossing, which is consistent with the previous gene-based estimates ([92], [78]). The
diversity landscapes in both species follow the domain-like organization of chromosomes
with higher diversity levels in the peripheral regions of higher recombination and low
diversity in the “central” domains of lower recombination.

In the absence of selection, evolutionary simulations show that nonuniform recombination
rates by themselves cannot generate the structure of diversity observed in the empirical
data, although it does of course have a very large impact on the majority of population
genetic statistics (Fig 6 and S10 Fig, neutral scenarios in the top rows compared to Fig 4
and S2 Fig). In general, these patterns agree with previous predictions expected for a
number of these statistics (see Table 1 in [43]).

**Interaction of selection and recombination**

Genetic variation and recombination rate tend to be correlated due to the reduction of
diversity of linked sites both by positive and background selection ([36], [138], [139], [38],
[140], [141], [4]). Positive correlations of meiotic recombination rate and nucleotide
diversity have been shown for many species ([142], [143], [144], [145], [108], [146], [147],
[148], [149]), however, this pattern is not universal and can be influenced by a variety of
factors ([107], [150]). Selfing generates a distinct reduction in effective recombination,
thereby reducing the efficiency of both background and positive selection ([151], [152],
[32], [39], [122], [42]). Both *C. elegans* and *C. remanei* display higher diversity in regions
of higher recombination, which may be a signature of linked selection. This is consistent with prior research showing that Caenorhabditis nematodes exhibit substantial effects of background ([144], [147], [153], [154]), positive selection ([66]), and perhaps balancing selection on some loci ([125], [98]). Our observation that exons and introns show very strong differences in diversity in C. remanei whereas they are very similar in C. elegans, provides strong evidence that selection is much more efficient in C. remanei than C. elegans, almost certainly because reduced effective recombination rates in C. elegans.

The genome organization across Caenorhabditis as a whole appears to be likely the result of prolonged continuous selection, as chromosomes within all species are compact, densely packed with genes, and have similar patterns of recombination, and chromatin activity ([77]). They also tend to lack repetitive elements in the regions of lower recombination, which are likely to have been removed by background selection ([133]). Similarly, the central gene-dense regions in domains of lower recombination also tend to be more conserved. For example, within C. remanei, reconstructing ancestral states using C. latens genome as an outgroup reveals that the central chromosome domains are twice as conserved as the arms, a pattern also observed in C. elegans ([132]).

But exactly how do recombination and selection interact to generate these patterns? To address this question, we performed evolutionary simulations using the Caenorhabditis domain-like recombination landscape under conditions of different selfing rates, patterns of deleterious and beneficial mutations, as well as mutations under balancing selection. By including selection, we were able to mimic the shape of some of the diversity statistics found in empirical data (Fig 2, S2 and Fig 6, 10, S11, S12). Importantly for the interpretation of diversity within C. elegans, even a small proportion of outcrossing (>1/N) was enough to start observing the effects of linked selection, as is predicted by theory ([121], [122], [120]). So, at first glance, it would appear that we can do a very good job predicting patterns of genomic diversity within these species using the right balance of natural selection integrated across domains that differ dramatically in recombination structure. However, diversity is not shaped solely by the forces that either restructure or remove it from the population, but also by mutational forces that introduce it into the population in the first place.

**Mutation landscapes**

Mutation is the initial source of genetic variation, and the mutation rate can fluctuate along the chromosome due to such factors as chromatin accessibility, methylation, transcription activity, recombination rate, genomic context, replication and reparation timing ([155], [156], [157], [158], [159], [160], [161], [162], [163], [164]). Nonuniform mutation
landscapes have been observed in various species ([165], [166], [167], [168]), including C. elegans ([103], [106]), where the mutation rate on the arms is 1.2-1.6 higher than in the central regions. For these nematodes, these differences might be caused by the mutagenic properties of recombination itself, which varies strongly across the genome, or by a variety of other potential factors. To explore potential variation in mutation rates across the genome, we inferred the substitution spectrum for C. remanei using our inference of ancestral genomic states. This spectrum has the same dominant type of substitution, C>T|G>A, as C. elegans ([103], [169], [170]), and these are the most common polymorphic types of biallelic substitution both in C. elegans and C. remanei samples (resp. 58.4% and 60% from all types of biallelic substitutions). The fraction of C->T|G->A, A->T|T->A, and C->G|G->C substitutions are different within regions of higher and lower recombination in C. remanei (S3 Fig). The similarity of both chromosome organization and substitution spectrum of C. remanei and C. elegans implies that the mutation rate in C. remanei might be greater in regions of high recombination. However, direct mutation accumulation experiments and more detailed analysis of mutational signatures and subtypes are required to fully describe the features of the mutation landscape of C. remanei. Similarly, we know very little about the degree of gene conversion within Caenorhabditis species, which can lead to an increase in homozygosity in the regions of higher recombination ([127]). An indirect signature of this process is higher GC content in the regions of higher recombination, which we do indeed observe in these Caenorhabditis species ([132], [133], [77]). However, since we currently have no basis for parameterizing these potential effects, the potential for gene conversion was not included in our simulations.

Patterns of genomic diversity with varying patterns of mutational input are uncomfortably similar to those produced by an interaction of selection recombination alone. It naturally makes sense that regions with more mutation would also have more standing variation. If higher mutation rates are correlated with higher recombination rates, then at least at a superficial level, it would seem to be very difficult to distinguish among these evolutionary forces (Fig 6 and Figures S10, S11, S12). Our evolutionary simulations showed that variation in mutation rate along the genome affects most diversity statistics that we used in this study, except Tajima's D, \( \omega \), ZnS, \( F_is \), and TMRCA. In reality, a complex combination of evolutionary forces influences the genetic variability in these nematodes, which suggests that a deeper understanding of the balance of these forces requires more subtle ways of distinguishing among them, as we attempt to do with our classification analysis below.

**History and structure of populations**
Organisms live in dynamic environments that change in space and time, which naturally has the potential to dramatically affect population densities and therefore the context for evolutionary change. \textit{C. elegans} populations locally undergo phases of exponential growth in localized areas of vegetative decay, followed by dispersal to habitats with new resources ([171], [124], [68]). Consistent with this, its global population structure suggests metapopulation dynamics of frequent local extinctions followed by recolonization ([172], [94], [111]).

The structure of the local collection of \textit{C. elegans} used here, which is part of a much larger dataset from Hawaii ([83]), shows both multiple divergent lineages and resampling of a few closely related individuals, consistent with the emerging metapopulation paradigm. And indeed, the inferred demographic history of this population suggests massive reduction and fluctuations in population size over time (Fig 5). While the ancient pattern of population history and size is comparable to the previously reported dynamics in \textit{C. elegans} described by Thomas et al. [73], these demographic histories cannot be directly compared on the recent time scale since we employed genomic data from nematodes isolated from one location, whereas they used pseudo-diploids from a "global" sample.

The global population structure of \textit{C. remanei} similarly implies extensive migration across its range ([82], [173]), but local dynamics are still poorly understood. In the Toronto sample of \textit{C. remanei}, we observed elevated inbreeding coefficient and background linkage disequilibrium along the genome, which is likely the result of within-family mating or similar demographic processes in recent generations. Moving back in time, the inferred evolutionary population history of this population displays major fluctuations in the population size of up to two orders of magnitude (Fig 5 and S8A Fig). Even so, the inferred sizes are larger than those predicted for \textit{C. elegans}. One consequence of these estimates is that it means that the total size and overall scale of population dynamics make it particularly difficult to implement this specific demography in our simulations due to excessive computational demands. Nevertheless, our simulations show that continuous exponential population growth had much stronger effects on most diversity statistics than fluctuations in the population size. Similarly, a recent bottleneck in outcrossing populations, which seems to be the case for the \textit{C. remanei} population, can have a significant impact on most of the diversity statistics used in our study ("outcrossing inbreeding" columns in Fig 6, and Figures S10, S11, S12). Overall, however, these demographic effects are minor relative to persistent differences in the mating system.

A potential explanation for the pattern of diversity within \textit{C. elegans} is that the pattern of polymorphism across the genome is a residual echo of ancestral outcrossing, even in the expected homogenizing effects of self-reproduction. In particular, \textit{C. elegans} is predicted to have switched to selfing reproduction within the last \(\sim 4\) mya ([47]). To test this
hypothesis, we simulated populations exhibiting a discrete change in the mode of reproduction and then tracked how the diversity decays over time. We find that polymorphism from the outcrossing ancestor is predicted to decay very rapidly, so it is very unlikely that the pattern of diversity that we observe in *C. elegans* is the ghost of past outcrossing and instead that it is maintained by ongoing low levels of outcrossing within extant populations or as a result of a complex interaction of selection and metapopulation dynamics.

**Evolutionary simulations and approach to classify evolutionary scenarios**

No complete theoretical framework exists to describe how all of the above factors — selection, drift, mutation, genome organization, mating system, and population history — interact to shape diversity at a genomic scale, despite much theory ([174], [42], [120], [122], [43], [175]). Forward-time evolutionary simulation provides an alternative approach to test hypotheses for complex scenarios designed specifically for the species of interest ([8], [9], [176]). Our simulated diversity landscapes compared with the empirical patterns in *C. elegans* and *C. remanei* confirmed the general conclusion that differences in effective population size are responsible for generating the scale and magnitude of genetic variation, whereas effective recombination and selection coupled with mutation and genome organization, shape the distribution of that diversity across the genome (the genomic landscape of diversity). Thus, changing modes of mating across *Caenorhabditis* has had a profound effect on both the scale and shape of the diversity landscape.

We successfully used neural networks to classify various potential evolutionary scenarios using known patterns generated by our forward-time evolutionary simulations that varied patterns of selection, selfing rate, and mutation landscape. Here, perhaps not surprisingly, selfing rate led to the most effective means of classifying potential outcomes. Overall, however, we were not particularly successful in classifying the specific empirical patterns observed in these species (Fig 8). Specifically, whereas patterns for *C. elegans* worked well from the viewpoint of our simulation results, those from *C. remanei* did not, as its patterns of genomic variation were frequently classified as being consistent with selfing rather than outcrossing. Particularly important here seems to be the fact that even small changes in the ratios of some statistics between domains can alter the predicted structure (see an example in S18 Fig; also, Figures S10, S11, S12).

Several factors could make our classification scheme particularly sensitive to these issues. First and foremost, we could not simulate all possible demographic histories of the population, because simulating populations of equivalent size to natural populations, as well as including large-scale changes in population size, turns out to be computationally infeasible. Nevertheless, it is clear that these effects can be substantial,
as additional simulations of both recent size fluctuation and moderate exponential population growth showed that recent changes in the population size affect most of the statistics used for classification (S13 Fig). Further, even subtle factors, such as the quality of genomic coverage and bioinformatic assumptions, can lead to significant differences among the diversity statistics (S19 Fig). Long-read data may help close the gap in the coverage and improve the mapping quality for species with such high diversity as we see with these outcrossing nematodes. Lastly, it is possible that the precision of the classification could be improved by including additional genomic drivers such as gene conversion. For *C. remanei*, some of these challenges might be overcome by including broader sampling from across the entire species range so that local demographic differences do not unduly influence the core species-wide diversity patterns generated by conserved genomic processes such as patterns of recombination and mutation. It is equally likely that the sensitivity of this approach can be enhanced when we are able to add in additional complexities such as realistic demographic histories and local spatial and temporal dynamics to the simulations.

**Conclusion**

*Caenorhabditis* nematodes provide a useful model to study evolutionary consequences of selfing, as they have highly divergent mating systems while maintaining an overall similarity in genome organization. We demonstrated the recombination landscape of outcrossing *C. remanei* to be similar to *C. elegans*, with extensive domains of lower recombination on the central parts of chromosomes that predict genomic landscape of the diversity in both species. The scale of genetic polymorphism within selfing *C. elegans* and outcrossing *C. remanei*, however, are dramatically different because of large differences in effective population size and demographic history. Although we were able to classify evolutionary forces that shaped diversity in simulated populations, applying this approach to empirical data will require more comprehensive simulations of realistic biological scenarios and additional high-quality empirical genomic data. These findings support the emerging perspective that understanding patterns of variation at any particular site in the genome requires a global perspective of the forces that shape variation across the genome as a whole.

**Materials and methods**

**Genetic map for *Caenorhabditis remanei***
We constructed a genetic map for *C. remanei* from crosses of 2 inbred strains, PX506 and PX553. Initially, *C. remanei* isolates were derived from individuals living with terrestrial isopods (*Oniscidea*) taken from Koffler Scientific Reserve at Jokers Hill, King City, Toronto, Ontario, in October 2008 as described in [177]. Strain PX390 was created from one female mated with 3 males from an isopod. Strain PX393 is from one female and one male from an independent isopod. The strains were propagated for 2-3 generations before freezing. PX506 and PX553 are inbred strains generated from PX390 and PX393, respectively (the parental strain for PX506 was inadvertently specified as PX393 in [77]). To reduce residual heterozygosity, the lines were sib-mated for 28-30 generations before freezing. Nematodes were kept under the standard laboratory conditions according to Brenner [49].

The genetic map was constructed from 4 crosses of *C. remanei* strains PX506 and PX553, (2 crosses of ♀PX506 x ♂PX553 and 2 ♀PX553 x ♂PX506) using parental genotypes and 341 individually sequenced F2 worms, all females. Single L4 animals were digested in proteinase K, and the DNA content was linearly amplified with the phi-29 enzyme (GenomiPhiV3, GE life sciences), then normalized samples were processed for bestRAD sequencing [178] with EcoRI restriction site based adapters. Each multiplexed sample set was sequenced on the Illumina Hi-Seq 4000 platform with four lanes of 100bp paired-end reads (University of Oregon Sequencing Facility, Eugene, OR). Additionally, we sequenced the PX553 parental strain using the Nextera kit (Illumina) and Hi-Seq 4000 platform. The genome sequence of the PX506 strain was generated previously [77].

For parental strains, we checked the sequence quality of reads with FastQC v.0.11.5 [179] and MultiQC v.1.3 [180], trimmed and filtered reads with Skewer v.0.2.2 [181]. The filtered reads were mapped to the *C. remanei* genome (GCA_010183535.1 from the NCBI database) with BWA-MEM v.0.7.17 [182]. Then we filtered reads with SAMtools v.1.5 [183], removed duplicates by Picard tools v.2.17.6 [184], realigned indels, and called variants with GATK v.3.7 and v.4.1 [185]. Variants were filtered with standard GATK hard filters (see discussions in [186], [187] [188]), with only diallelic loci being used in the analysis. We also masked repetitive regions, as well as sites with too low or too high of coverage.

To process the bestRAD reads, we filtered reads without barcodes and flip forward and reverse reads when a barcode was found on the reverse read using Flip2BeRAD [189], demultiplexed reads with process_radtags from the Stacks package v.1.46 [190], followed by an additional adapter and quality trimming with Skewer. Then, we mapped the reads to the *C. remanei* reference genome (GCA_010183535.1) using bwa mem, marked duplicates and recalibrated alignment with variants from the parental, PX506 and PX553, strains with Picard and GATK BaseRecalibrator, filtered reads that do not cover the...
parental variants, have secondary alignments, or low mapping quality by SAMtools, and called variants with samtools mpileup. We generated the genetic map with Lep-Map3 [191], the order of markers was defined by their positions on the reference genome. The recombination rate per Mb was estimated in R, the boundaries of low and high recombination domains were determined with the pricewise regression by the segmented package [192] in R. For more details on this part of the analysis, see the scripts at https://github.com/phillips-lab/CR_CE_popgen/genetic_map/.

C. remanei and C. elegans population genomic data

To study the genomic pattern of diversity in outcrossing nematodes, we sequenced 14 wild individuals of C. remanei. Isopods, a phoretic host carrier of C. remanei [193], were collected at the same station as the strains used for the genetic map, the Koffler Scientific Reserve in Ontario, Canada, in September 2013, sacrificed within a few hours following collection after having been placed on agar plates seeded with Escherichia coli OP50. From each of 14 single C. remanei individuals isolated the next day from these samples (S3 Table), genomic DNA was directly amplified using the Repli-G kit (Qiagen), and then sequenced with Illumina HiSeq from TruSeq gDNA libraries by GenomeQuebec. One pair of C. remanei individuals derived from a shared isopod host (NS50-1, NS50-2), whereas all other individuals were isolated from different isopods.

To compare the population diversity patterns of this outcrossing species with a selfing species using similar approaches, we re-analyzed genomic sequences of 28 wild isolates of C. elegans collected at one location on the Big Island, Hawaii, from Crombie et al. [83]. For more details on the C. elegans sample, see Source data 1 from [83]. Sample IDs and the NCBI Sequence Read Archive accession numbers for C. elegans and C. remanei are in S1 Table.

Variant calling, diversity statistics, and demography

We filtered and mapped reads to the C. elegans genome (project PRJNA13758, from WormBase version WS245) or the C. remanei genome (GCA_010183535.1 from the NCBI database) as described for the genetic map parental strains above. Variants were filtered with the standard GATK hard, only diallelic loci were used in the analysis. Additionally, we masked some genomic regions, such as indels with 10bp flanking regions with GATK and BEDTools v.2.25 [194]; repetitive regions using the masked versions of genomes and a script to extract them from [195]; and regions with too low or high
mappability (<5x or >100x coverage in half of the individuals), all masks were combined by BEDTools merge.

We estimated 12 population diversity statistics using diploS/HIC fvecVcf [196] with two minor modifications: allowing not to normalize statistics, and to have only one sub-window in a window (see diploSHIC_note.txt on the project GitHub), and, additionally, the β-statistic using BetaScan [197] for 100 kb windows in *C. elegans* and *C. remanei* samples. To estimate the β-statistic for empirical data, we first applied an imputation with beagle v.5.0 [198] and a data format conversion by glactools [199]. Nucleotide diversity within introns and exons was estimated using these features from corresponding genome annotations, VCFtools [200], and BEDTools. We compared nucleotide diversity between domains of recombination and different gene features using Cohen’s d from the lsr package [201] and the Fisher-Pitman Permutation Test (Z) from the coin package [202] in R [203].

Demographic history was inferred by SMC++ v1.15.1 [113] for 100 bootstrapped replicates of 8 individuals using data from chromosomes I, III, IV for *C. elegans* and I, II, III, V for *C. remanei*. Additionally, we performed this analysis for all chromosomes separately. The genomic regions with unfit mappability or repeats were masked as described above. We used the mutation rate of 2.3e-09 base substitution per generation [106], and, for *C. elegans*, we rescaled this mutation rate by 0.5 and, later, the obtained generation time by 2 due to selfing (as discussed in [204]). To compare estimates of the recent population sizes from SMC++, as well as with previous estimates within *C. elegans*, we also calculated effective population size from nucleotide diversity data using the Watterson estimator [205] assuming neutrality and complete selfing for *C. elegans* (see [32] and [206], Ne=N/2). The recent demographic histories based on the linkage disequilibrium information were inferred by GONE ([112]), using autosomes and recombination rate retrieved from the genetic maps of *C. remanei* and *C. elegans* (Table 1; and Table 1 from [65]); the population size was averaged by taking the geometric mean across 100 replicates. Generation time was scaled by 2, and population size by 0.5 for *C. elegans*, analogous to the approach outlined above.

The inbreeding coefficient (Fis) along the genome was estimated for sites with a minor allele frequency of more than 0.05 using VCFtools v.0.1.15, BEDtools v.2.25, HTSlib v.1.6, plink v1.90b4.6 [207], and popStats v1.0.0 [208]. The sample structure was visualized in the two-dimension space using latent coordinates from popVAE [209]. The effective recombination rate along the genomes was inferred by ReLERNN [176], using one of the replicates of reconstructed demography from SMC++ and (x0.5) mutation and (x2) time scaling for the selfing species. Linkage disequilibrium within and across chromosomes and the LD decay were estimated with plink using all sites for *C. elegans*.
sample, and every 10th site for *C. remanei* for all LD calculations except the fine-resolution LD where we used all sites.

Figures were plotted via R packages dichromat ([210]), ggplot2 [211], gridExtra [212], ggpubr ([213]), magick ([214]), and also boot ([215]), coin ([216]), lsr ([217]), reshape2 ([218]), scales ([219]) to estimate summary statistics. The scripts for variant calling and masking, diversity statistics estimation, demography reconstructions, populations structure, effective recombination inference, LD decay, and corresponding figures and statistics are in https://github.com/phillips-lab/CR_CE_popgen/diversity_stats/.

**Reconstruction of the ancestral states of *C. remanei***

We reconstructed the ancestral states of the *C. remanei* and *C. latens* reference genomes. We used genomes of four strains of *C. remanei*, PX506, PX356 (GCA_001643735.2), PB4641 (GCA_000149515.1), PX439 (GCA_002259225.1) and *C. latens* (GCA_002259235.1). Preliminary topology was obtained with progressiveMauve [220] with 1 Mb regions from each of 6 chromosomes. Next, the genomes were masked based on their mappability by GenMap [221], and aligned by Progressive Cactus [222] with the following species tree topology: (((*C. remanei* PX439), *C. remanei* PX506), *C. latens*). The ancestral states were re-estimated by ancestorsML tool in the HAL tools [223].

Then for each chromosome, we calculated the fractions of sites with ancestral states, GC content in the ancestor, various types of substitutions per 100 kb windows, and extracted ancestral states positions polymorphic in the *C. remanei* population by HAL tools, BEDtools, and bash. We used Relate [114] to infer the history of the *C. remanei* population based on the ancestral states and the recombination map described above. We also examined demographic tests of the signal of positive selection along the genome, adjusting reported p-values using the harmonic mean p-value within 3Mb windows (2 Mb overlap) from harmonicmeanp [224] R package. The time to the most recent common ancestor (TMRCA), the relative TMRCA half-life (RTH, following [225]), and terminal branch lengths were assessed by phangorn [226] and phytools [227] R packages from Relate generated trees. Statistical analysis and visualization was made with coin, data.table ([228]), dplyr ([229]), ggplot2, gridExtra, ggpubr, lsr, pals ([230]), and scales packages in R. All scripts for the ancestral reconstruction, demographic inference, and related analyses available at https://github.com/phillips-lab/CR_CE_popgen/ancestral/
Evolutionary simulations

To understand how various factors affect the genomic landscape of diversity, we run forward-time individual-based evolutionary simulations in SLiM v.3.3 ([8], [9]) using the tree-sequence format ([10]). We performed three types of simulation: 1) effects of the selfing rate, selection, and mutation landscape; 2) the decay of the ancestral diversity; 3) effects of rapid changes in population size on estimated statistics. All simulations had a population size of 5,000 individuals, 1 chromosome of 3 Mb in size with three recombination domains: left and right arms (1 Mb) with a high recombination rate (2.5e-7), and a 1 Mb central domain of low recombination (1e-9) to mimic the recombination domains of lower and higher recombination in nematodes.

In the first type of simulations, we changed selfing, mutation landscape, and selection regime. The selfing rate was 0% (outcrossers), 90%, 98%, 99.9%, or 100% (selfers). The mutation landscape was uniform with the mutation rate of 2e-8 in all domains of recombination. While this is an order of magnitude higher than actual mutation rate estimates, higher rates greatly facilitate the simulation process and, since the emphasis is on the relative values of mutation on arms and centers, this difference should not affect the normalized statistics used above. We used several domain-specific patterns of mutation rate differences: a uniform landscape with no difference in the mutation rate on the arms and the central domain (denoted as 1-1-1), 15% more mutations on the arms than on the centers (1.15-1-1.15), 50% more (1.5-1-1.5), or 2 times more mutations on the arms (2-1-2), the 1x mutation rate was the same in all simulations. Slightly elevated mutation rates within regions of higher recombination have been shown for worms previously [32], [49]. We considered four main types of selection regimes: only neutral mutations; neutral and 10% deleterious mutations; neutral, 10% deleterious, and 1% beneficial mutations; neutral, 10% deleterious, and 1% balancing mutations. Non-neutral simulations utilized various distributions for selection, and simulations with balancing selection used only one of these distributions (see details in S4 Table). Distributions of dominance were different for deleterious and beneficial mutations, with a shift towards recessive for the former and more additive coefficients for the latter (see S20 Fig and the SLiM scripts). All simulations were run for 50,000 generations (10N, burn-in).

The second type of simulations was designed to explore the effects of a switch from ancestral outcrossing to selfing. We ran neutral and non-neutral simulations (see details in S4 Table) in which all populations were outcrossing during burn-in and then subsequently changed to either 98 or 100% selfing. To observe the decay of ancestral diversity we saved results every 5,000 generations for 30,000 (6N) generations and repeated each scenario 50 times.
The aim of the third type of simulations was to reveal the consequences of rapid change in population size on the diversity statistics used in our analyses. These simulations utilized only neutral mutations but allowed different selfing rates and mutational landscapes. First, we performed simulations with 100 generations of exponential growth of 3% following burn-in, generating a rapid population size increase from 5K to about 100K. Second, we investigated the effects of fluctuations in population size following burn-in, by setting the population size to 15,000 and then back to 5,000 for 10 generations and so on 5 iterations. For the third type of scenario with exponential growth or fluctuations of the population size, we compared diversity statistics for the same populations at the burn-in generation and at the end of the simulation for 50 replicates.

We added neutral mutations, “recapitated” trees, and converted tree sequences from SLiM simulations to the VCF file format with 100 individuals using tskit v.0.3.4 ([231]), msprime v.0.6.1 ([232]), pyslim ([233]), and estimated diversity statistics with diploS/HIC and BetaScan as for empirical data but for 40kb windows with and without normalization. We also estimated the divergence from the ancestral genome using bash and BEDtools, Fis statistics via popStats, and calculated tree heights (TMRCA) from tree-sequences using python modules argparse ([234]), msprime, statistics ([239]), and pyslim.

We plotted and analyzed simulation results using packages coin, corplot, dichromat, dplyr, ggplot2, gridExtra, ggrepur, and lsr packages in R. SLiM scripts for simulations, and related R scripts for visualization and analysis are in https://github.com/phillips-lab/CR_CE_popgen/simulations/.

Classification of empirical data

In order to classify mutation rate, selfing rate, and type of selection on empirical data we trained a convolutional neural network (CNN) on a variety of features from the simulated chromosomes. Each simulated chromosome was divided into 75 sub-windows and 13 population diversity statistics were calculated within each sub-window, resulting in 13 by 75-pixel images used as input features for the CNN. The targets for the network consisted of one-hot encodings on each combination of the three evolutionary scenarios.

PCA and saliency plots were employed to visualize the similarity between simulated and empirical features. We projected the empirical features to PC space fitted to the simulated features in order to see how close the simulations resemble the empirical features. Saliency plots allow us to glimpse the features through the lens of the CNN. That is, we can see how sensitive the network classification is to perturbations of the summary statistics in each sub-window.
The CNN architecture was composed of four convolutional layers and four dense, fully connected layers. To prevent overfitting, dropout was used in between all but the final two layers. ReLU activations were used on all but the final layer, which had linear activation to improve numerical stability for the categorical cross-entropy loss functional. Weight learning was controlled by the Adam optimization routine with an initial learning rate of 1.5e-5. To reduce overfitting, we stopped training after the validation loss improved no more than 1e-7 for 100 epochs and saved the weights that resulted in the highest validation accuracy. The CNN was implemented using the Keras module within TensorFlow ([235]). The details of the analysis are at https://github.com/phillips-lab/CR_CE_popgen/cnn_model/.

Acknowledgments

We thank Rose Reynolds, Timothy Ahearne, Chadwick Smith, Scott Scholz for C. remanei mapping strain generation, Larry Meng for DNA isolation of C. remanei worms from the Koffler Scientific Reserve used for population genomic sequencing. Many thanks go to Sean O’rourke and members of Mike Miller’s lab at UC Davis for help with bestRAD sequencing, as well as to Gavin Woodruff, Murillo Rodrigues, Peter Ralph, Andy Kern, and other former and current members of Phillips lab and Kern-Ralph co-lab at the University of Oregon for comments and discussions. We thank the Research Advanced Computing Services team for assistance with the computing cluster Talapas at the University of Oregon and the Genomic Core Facility at the University of Oregon for assistance with library preparation.

References


Incompatibility and the Evolution of Recombination. PLoS Genet. 2011;7:
e1002174. doi:10.1371/journal.pgen.1002174

77. Teterina AA, Willis JH, Phillips PC. Chromosome-Level Assembly of the
Caenorhabditis remanei Genome Reveals Conserved Patterns of Nematode
doi:10.1534/genetics.119.303018

78. Jovelin R, Ajie BC, Phillips PC. Molecular evolution and quantitative variation for

79. Cutter AD, Charlesworth B. Selection Intensity on Preferred Codons Correlates
with Overall Codon Usage Bias in Caenorhabditis remanei. Current Biology.

80. Cutter AD, Baird SE, Charlesworth D. High Nucleotide Polymorphism and Rapid
Decay of Linkage Disequilibrium in Wild Populations of Caenorhabditis remanei.

81. Jovelin R, Dunham JP, Sung FS, Phillips PC. High Nucleotide Divergence in
Developmental Regulatory Genes Contrasts With the Structural Elements of
doi:10.1534/genetics.107.082651

1269. doi:10.1534/genetics.112.140418

sampling of Hawaiian Caenorbhitis elegans reveals high genetic diversity and

84. Wagner CR, Kuervers L, Baillie D, Yanowitz JL. xnd-1 Regulates the Global
doi:10.1038/nature09429

85. Meneely PM, McGovern OL, Heinis FI, Yanowitz JL. Crossover Distribution and
Frequency Are Regulated by him-5 in Caenorhabditis elegans. Genetics. 2012;190: 
1251–1266. doi:10.1534/genetics.111.137463

86. Yu Z, Kim Y, Dernburg AF. Meiotic Recombination and the Crossover Assurance
doi:10.1016/j.semcdb.2016.03.014

87. Zetka MC, Rose AM. Sex-Related Differences in Crossing over in Caenorhabditis

88. Meneely PM, Farago AF, Kauffman TM. Crossover distribution and high
interference for both the X chromosome and an autosome during oogenesis and
89. Lim JGY, Stine RRW, Yanowitz JL. Domain-Specific Regulation of Recombination in Caenorhabditis elegans in Response to Temperature, Age and Sex. Genetics. 2008;180: 715–726. doi:10.1534/genetics.108.090142


