# Balancing selection maintains ancient genetic diversity in *C. elegans*

Daehan Lee[*,1,10], Stefan Zdraljevic[*,1,2,11,12], Lewis Stevens[*,1], Ye Wang[1], Robyn E. Tanny[1], Timothy A. Crombie[1], Daniel E. Cook[1,13], Amy K. Webster[3,4], Rojin Chirakar[3], L. Ryan Baugh[3,5], Mark G. Sterken[6], Christian Braendle[7], Marie-Anne Félix[8], Matthew V. Rockman[9], Erik C. Andersen[1]

1. Department of Molecular Biosciences, Northwestern University, Evanston, IL 60208, USA
2. Interdisciplinary Biological Sciences Program, Northwestern University, Evanston, IL 60208, USA
3. Department of Biology, Duke University, Durham, NC, USA
4. University Program in Genetics and Genomics, Duke University, Durham, NC, USA
5. Center for Genomic and Computational Biology, Duke University, Durham, NC, USA
6. Laboratory of Nematology, Wageningen University and Research, 6708PB, The Netherlands
7. Université Côte d'Azur, CNRS, Inserm, IBV, France, 06100 Nice, France
8. Institut de Biologie de l'Ecole Normale Supérieure, Centre National de la Recherche Scientifique, INSERM, École Normale Supérieure, Paris Sciences et Lettres, Paris, France
9. Center for Genomics and Systems Biology, Department of Biology, New York University, New York 10003, USA
10. (Present address) Center for Integrative Genomics, University of Lausanne, CH-1015 Lausanne, Switzerland
11. (Present address) Department of Human Genetics, University of California, Los Angeles, CA 90095, USA
12. (Present address) Howard Hughes Medical Institute, University of California, Los Angeles, CA 90095, USA
13. (Present address) The Francis Crick Institute, London NW1 1ST, UK

* These authors contributed equally to this work.
Correspondence should be addressed to E.C.A (erik.andersen@northwestern.edu).

**Erik C. Andersen**
Associate Professor of Molecular Biosciences
Northwestern University
Evanston, IL 60208, USA
Tel: (847) 467-4382
Email: erik.andersen@northwestern.edu

Daehan Lee, daehan.lee@unil.ch, ORCID 0000-0002-0546-8484
Stefan Zdraljevic, szdraljevic@mednet.ucla.edu, ORCID 0000-0003-2883-4616
Lewis Stevens, lewis.stevens07@gmail.com, ORCID 0000-0002-6075-8273
Ye Wang, ye.wang@northwestern.edu, ORCID 0000-0002-5423-6196
Robyn E. Tanny, robyn.tanny@northwestern.edu, ORCID 0000-0002-0611-3909
Timothy A. Crombie, tcrombie@northwestern.edu, ORCID 0000-0002-5645-4154
Daniel E. Cook, danielecook@gmail.com, ORCID 0000-0003-3347-562X
Amy K. Webster, amy.k.webster@duke.edu, ORCID 0000-0003-4302-8102
Rojin Chirakar, rc226@duke.edu
L. Ryan Baugh, ryan.baugh@duke.edu, ORCID: 0000-0003-2148-5492
Mark G. Sterken, mark.sterken@wur.nl, ORCID 0000-0001-7119-6213
Christian Braendle, braendle@unice.fr, ORCID 0000-0003-0203-4581
Marie-Anne Félix, felix@biologie.ens.fr
Matthew V. Rockman, mrockman@nyu.edu, ORCID 0000-0001-6492-8906
Erik C. Andersen, erik.andersen@northwestern.edu, ORCID 0000-0003-0229-9651

**Summary paragraph**

The mating system of a species profoundly influences its evolutionary trajectory[1–3]. Across diverse taxa, selfing species have evolved independently from outcrossing species thousands of times[4]. The transition from outcrossing to selfing significantly decreases the effective population size, effective recombination rate, and heterozygosity within a species[5]. These changes lead to a reduction in the genetic diversity, and therefore adaptive potential, by intensifying the effects of random genetic drift and linked selection[6,7]. Selfing has evolved at least three times independently in the nematode genus *Caenorhabditis*[8], including in the model organism *Caenorhabditis elegans*, and all three selfing species show substantially reduced genetic diversity relative to outcrossing species[8,9]. Selfing and outcrossing *Caenorhabditis* species are often found in the same niches, but we still do not know how selfing species with limited genetic diversity can adapt to and inhabit these same diverse environments. Here, we discovered previously uncharacterized levels and patterns of genetic diversity by examining the whole-genome sequences from 609 wild *C. elegans* strains isolated worldwide. We found that genetic variation is concentrated in punctuated hyper-divergent regions that cover 20% of the *C. elegans* reference genome. These regions are enriched in environmental response genes that mediate sensory perception, pathogen response, and xenobiotic stress. Population genomic evidence suggests that these regions have been maintained by balancing selection. Using long-read genome assemblies for 15 wild isolates, we found that hyper-divergent haplotypes contain unique sets of genes and show levels of divergence comparable to that found between *Caenorhabditis* species that diverged millions of years ago. Taken together, these results suggest that ancient genetic diversity present in the outcrossing ancestor of *C. elegans* has been maintained by long-term balancing selection since the evolution of selfing. These results provide an example for how species can avoid the evolutionary “dead end” associated with selfing by maintaining ancestral genetic diversity.

## Global distribution of *C. elegans* genetic diversity

To understand the evolutionary history of the *C. elegans* species, we examined whole-genome sequence data of 609 wild strains isolated from six continents and several oceanic islands (Fig. 1a, Supplementary Table 1). We identified 328 distinct genome-wide haplotypes (henceforth, referred to as isotypes) (Methods). We aligned sequence reads from all isotypes to the N2 reference genome[10] and characterized genetic variation across the species, including 2,431,645 single nucleotide variants (SNVs) and 845,797 insertions and deletions (indels, ≤ 50 bp). We used these variant data to identify the most highly divergent isotypes, which were isolated exclusively from the Pacific region, including Hawaii, New Zealand, and the Pacific coast of North America (Fig. 1a,b, Supplementary Fig. 1).

To further characterize the geographic population structure within the *C. elegans* species, we performed principal component analysis (PCA)[11] and found that most of the isotypes (326 isotypes) could be classified into three genetically distinct groups (Global, Hawaiian, and Pacific groups) (Fig. 1c-f, Supplementary Fig. 2). The largest Global group includes 93.9% (308) of all isotypes from six continents and oceanic islands (Fig. 1d). The Hawaiian group consists of eight isotypes from the Big Island of Hawaii, and the Pacific group includes ten isotypes from Hawaii, California, and New Zealand (Fig. 1e,f). Notably, Hawaii is the only location where the isotypes of all three groups have been found. For example, the Big Island harbors 25 isotypes from all three groups, whereas all 167 European isotypes belong only to the Global group (Fig. 1d-f). Furthermore, the two most divergent isotypes, XZ1516 and ECA701, which do not belong to any of the three groups, were sampled from Kauai, the oldest sampled Hawaiian island (Fig. 1c). This remarkable genetic diversity sampled from the Hawaiian Islands suggests that *C. elegans* likely originated from the Pacific region[12,13].

Next, we attempted to further subdivide the Global group using PCA. This group includes all non-Pacific isotypes and the majority of isotypes from the Pacific Rim. Although we found a weak genetic differentiation of Hawaiian, North American, and Atlantic isotypes from the rest of these isotypes, the Global group was not further clustered into distinct genetic groups (Supplementary Fig. 2, Extended Data Fig. 1a-c). Additionally, the Global group could not be separated into distinct geographic locations because many genetically similar isotypes have been sampled from different continents. By characterizing genomic regions predicted to be identical by descent, we found that the previously reported chromosome-scale

selective sweeps[13] contribute substantially to the genetic similarity we observed among geographically distant isotypes (Extended Data Fig. 1). For example, a large haplotype block on the center of chromosome V is shared by isotypes from six continents (Africa, Asia, Australia, Europe, North America, and South America). In addition to the Hawaiian isotypes that were reported to have avoided these selective sweeps[12], we found that the genomes of isotypes from Atlantic islands (*e.g.* Azores, Madeira, and São Tomé), in contrast to continental isotypes, show less evidence of the globally distributed haplotype that swept through the species (Extended Data Fig. 2, Supplementary Table 2). Taken together, these results suggest that recent selective sweeps could have occurred along with the out-of-Pacific expansion of *C. elegans*.

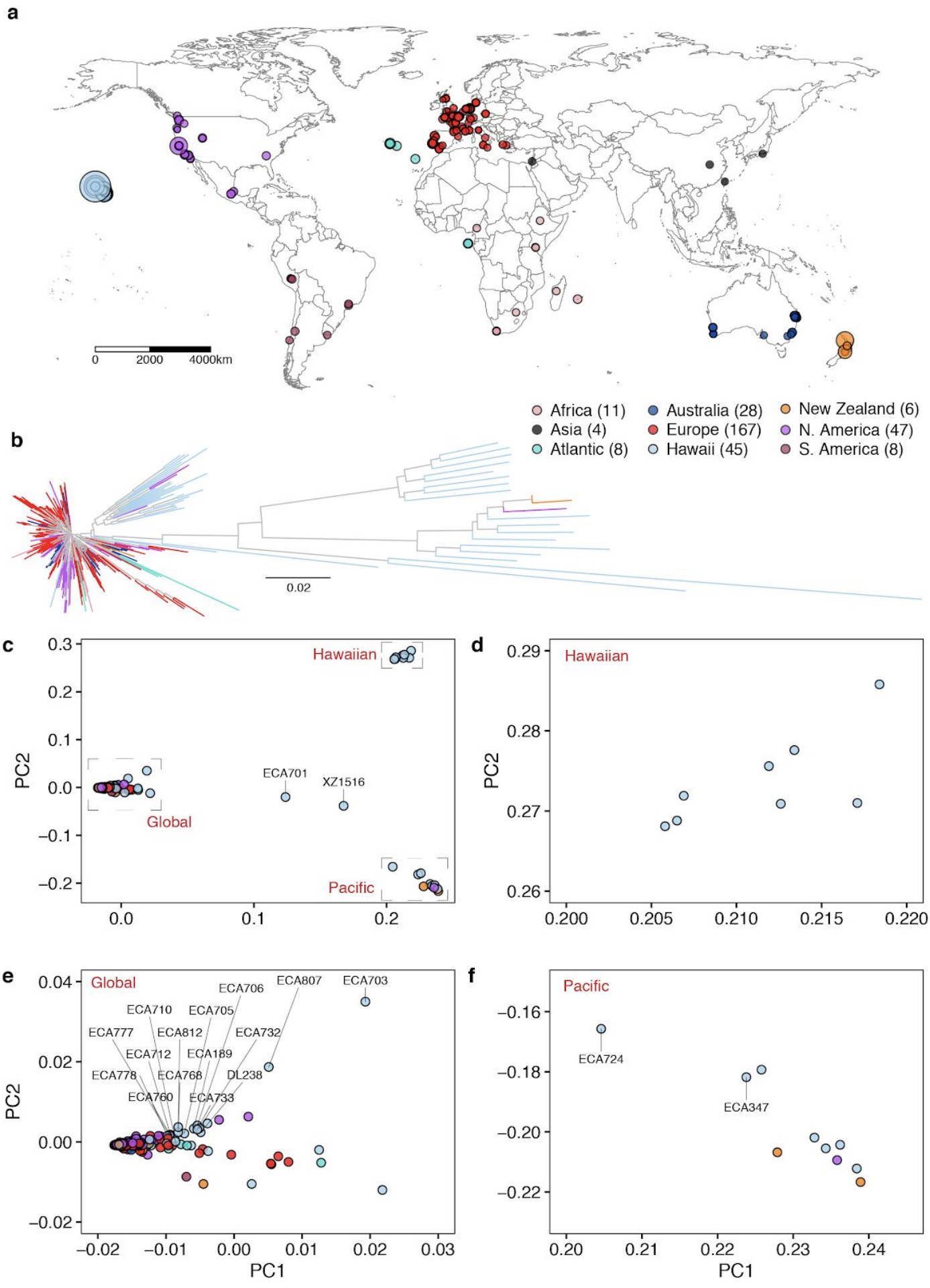


**Fig. 1 | Genetically divergent wild *C. elegans* strains were isolated from the Pacific region**
(a) The global distribution of 324 isotype reference strains is shown. Each circle corresponds to one of the 324 wild isotypes and is colored by the geographic origin. Four isotypes that do not have geographic information are not shown. The size of each circle corresponds to the number of non-reference homozygous alleles. The number of isotypes from each geographic origin is labeled in parentheses.
(b) A neighbour-joining tree of 328 *C. elegans* wild isotypes generated from 963,027 biallelic segregating sites is shown. The tips for four isotypes with unknown geographic origins are colored grey and the other 324 isotypes are colored the same as (a) by their geographic origins.
(c) Plots of the 328 isotypes according to their values for each of the two significant axes of variation, as determined by principal component analysis (PCA) of the genotype covariances. Each point is one of the 328 isotypes, which is colored by their geographic origins same as (b). Two divergent isotype, XZ1516 and ECA701, are labeled with grey lines.
(d-f) Zoomed in plots of the Hawaiian, Global, and Pacific groups of (c), respectively. (e,f) Isotypes from the Big Island of Hawaii are labeled with grey lines.

## Discovery of species-wide hyper-divergent genomic regions

Previous studies have shown that genetic variation is distributed non-randomly across each of the five *C. elegans* autosomes, with the chromosome arms harboring more genetic variation than chromosome centers[13]. This distribution is thought to be shaped by higher recombination rates on chromosome arms than centers, which alters the effects of background selection[14–16]. Consistent with previous studies[13], we observed 2.2-fold higher levels of nucleotide diversity ($\pi$) on chromosome arms compared to the centers (Welch's t-test, $P < 2.2 \times 10^{-16}$) (Extended Data Fig. 3, Supplementary Fig. 3). Furthermore, we found that variation is concentrated in punctuated regions of extremely high divergence that are marked by the higher-than-average density of small variants and large genomic spans where short sequence reads fail to align to the N2 reference genome (Extended Data Fig. 4a, Supplementary Fig. 4).

We sought to characterize the species- and genome-wide distributions of these regions (henceforth, referred to as hyper-divergent regions). To facilitate the accurate identification of these hyper-divergent regions across the *C. elegans* species, we generated high-quality genome assemblies using long-read sequencing data for 14 wild isotypes that span the species-wide diversity (Methods, Supplementary Table 3). We first aligned these assemblies, along with a previously published long-read assembly of the Hawaiian isotype CB4856[17], to the N2 reference genome. Next, we used these alignments to classify hyper-divergent regions across a range of parameter values (*e.g.* alignment coverage and sequence divergence). We performed a similar parameter-search procedure using short-read alignments of these 15 isotypes and identified a set of short-read parameters that maximizes the overlap between long- and

short-read hyper-divergent classification (Methods, Supplementary Fig. 5). To validate our approach, we compared the hyper-divergent genomic regions we identified in the CB4856 strain to what was previously reported in this strain (2.8 Mb)[18] and found that we detected a similar number and extent of these regions (3.2 Mb). Finally, we applied these optimized short-read hyper-divergent region classification parameters to the entire set of 327 non-reference isotypes (Extended Data Fig. 4b,c).

Across all these isotypes, we identified 366 non-overlapping regions that are hyper-divergent in at least one isotype as compared to the reference isotype N2 (Fig. 2, Supplementary Table 4, Supplementary Data 1)[18]. These regions range in size from 9 kb to 1.35 Mb (mean = 56 kb, median = 19 kb) (Supplementary Fig. 6) and cover approximately 20% of the *C. elegans* N2 reference genome (20.5 Mb). The majority of these regions (69%) are found on autosomal arms and contain 10.3-fold higher variant (SNVs/indels) densities than the non-divergent autosomal arm regions (16.6-fold more than the genome-wide average) (Supplementary Table 5, Supplementary Fig. 7, Methods), similar to the level of genetic diversity reported within outcrossing *Caenorhabditis* species[8,9]. Across all of the non-reference isotypes, we found substantial differences in the extent of the genome that was classified as hyper-divergent (0.06 - 11.6% of the genome) (Fig. 2c,d). The genome of the XZ1516 strain contains a striking 11.7 Mb of hyper-divergent regions. On average, 20.2% of the variation in a typical isotype is localized to hyper-divergent regions that span 1.9% of the reference genome (Fig. 2c, Supplementary Table 2).

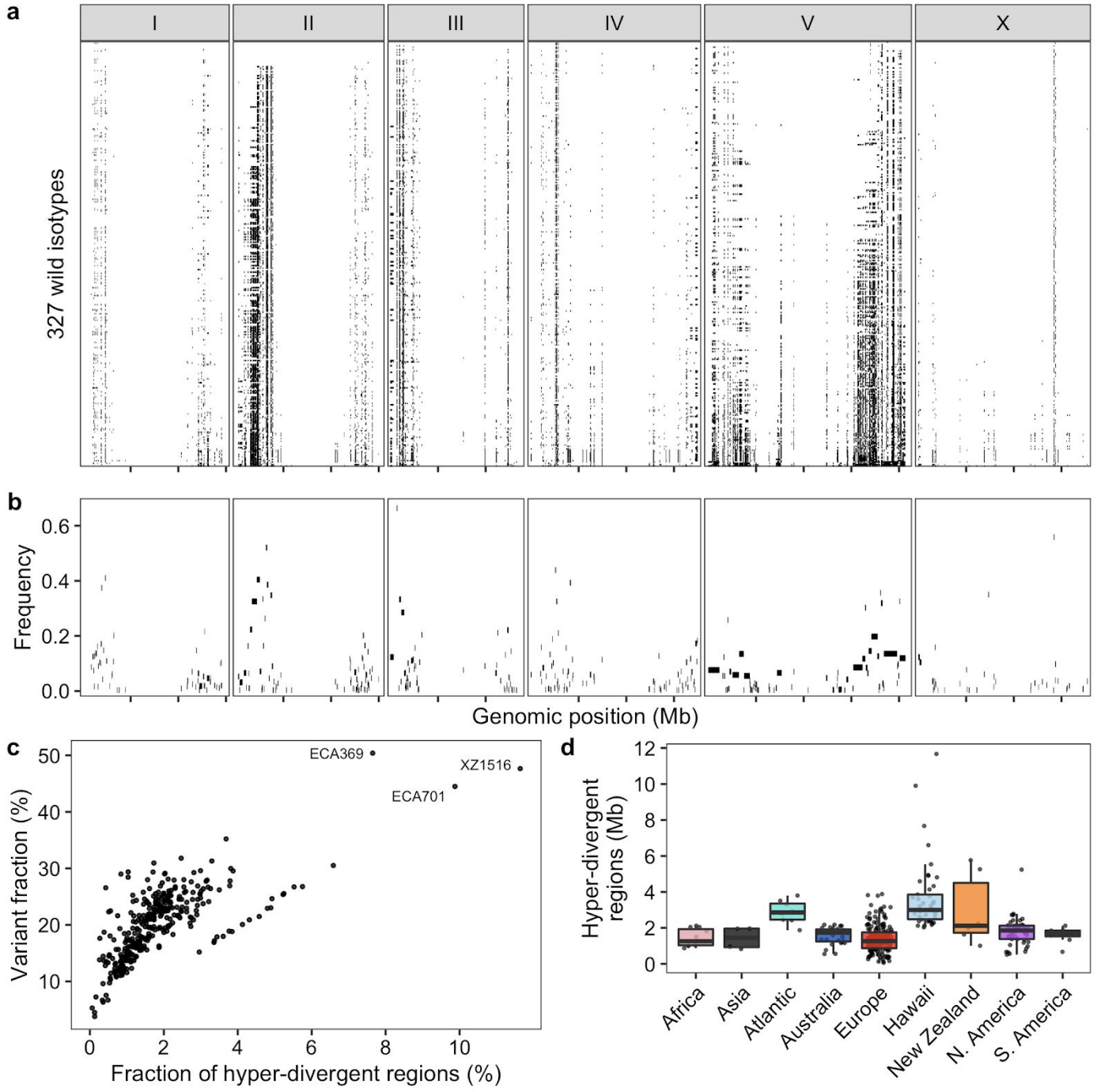



**Fig. 2 Punctuated hyper-divergent genomic regions are widespread across the *C. elegans* species**
(a) The genome-wide distribution of hyper-divergent regions across 327 non-reference wild *C. elegans* isotypes is shown. Each row represents one of the 327 isotypes, ordered by the total amount of genome covered by hyper-divergent regions (black). The genomic position in Mb is plotted on the x-axis, and each tick represents 5 Mb of the chromosome.
(b) The species-wide frequencies of 366 species-wide hyper-divergent regions flanked by non-divergent regions (Methods) are shown. Each rectangle corresponds to a block of species-wide hyper-divergent regions. The genomic position in Mb is plotted on the x-axis, and each tick represents 5 Mb of the chromosome. The average frequencies of hyper-divergent calls across 1 kb bins in each block across 327 non-reference wild isotypes are shown on the y-axis.
(c) A scatter plot showing the proportion of the N2 reference genome that is hyper-divergent in each isotype (x-axis) and the fraction of total variants in hyper-divergent regions of 327 non-reference isotypes (y-axis)

are shown. Each point corresponds to one of the 327 isotypes. The names of three isotypes with the largest genome-wide extents of hyper-divergent regions are shown.
(d) Tukey box plots of the total amount of genome covered by hyper-divergent regions are shown with data points plotted behind. Wild isotypes are grouped by their geographic origins. Each point corresponds to one of the 324 wild isotypes with a known geographic origin, and genome-wide fractions of hyper-divergent regions are shown on the y-axis. The horizontal line in the middle of the box is the median, and the box denotes the 25th to 75th quantiles of the data. The vertical line represents the 1.5x interquartile range.

**Maintenance of hyper-divergent haplotypes**

The species-wide distribution of hyper-divergent regions revealed that many non-reference isotypes are often hyper-divergent at the same genomic regions of the N2 reference genome (Fig. 2a,b, Fig. 3a). When compared to the reference genome, we found that these regions range from being divergent in a single isotype to divergent in 280 (85%) isotypes. Interestingly, we find that SNVs in hyper-divergent regions have a lower rate of linkage disequilibrium (LD) decay than SNVs within non-divergent genomic regions (Fig. 3b), suggesting that these regions are inherited as large haplotype blocks. Consistent with this interpretation, genomic regions that are classified as hyper-divergent among more than 5% of isotypes have elevated levels of Tajima's *D* (Fig. 3c), a commonly used statistic to identify regions under balancing selection[19,20], relative to non-divergent genomic regions. Together, these results suggest that wild isotypes frequently share hyper-divergent haplotypes that have been maintained in the *C. elegans* species by balancing selection.

Balancing selection can maintain genetic diversity that contributes to the adaptive potential of a population in the presence of environmental heterogeneity[21]. To investigate if hyper-divergent regions are functionally enriched for genes that enable *C. elegans* to thrive in diverse habitats, we performed gene-set enrichment analysis using two complementary approaches (Methods, Supplementary Data 2,3). Among the most significantly enriched gene classes were those related to sensory perception and xenobiotic stress (Fig. 3d, Supplementary Fig. 8, 9). We found that seven-transmembrane receptor class genes (*e.g.* G-protein coupled receptors, GPCRs) are overrepresented in hyper-divergent regions and that 54.9% (802/1461) of these genes are located in these regions. In addition to GPCRs, 48.2% (124/257) of genes that encode for C-type lectins, 53% (317/598) of the E3 ligase genes, and 86.8% (33/38) of the *pals* genes, which are involved in response to diverse pathogens[22–26], are found in hyper-divergent regions (Supplementary Fig. 10, Supplementary Data 2). In agreement with these enrichment results, we found that 66.8% (131/196) and 65.4% (85/130) of genes that are differentially expressed in the reference N2 isotype

upon exposure to the natural pathogens *Nematocida parisii*[27] or Orsay virus[28], respectively, are located in these regions[29]. Furthermore, we found that genes in hyper-divergent regions are more strongly induced than genes in non-divergent regions of the genome (Welch's t-test, *p*-value = 0.00606, for *N. parisii* and *p*-value = 0.0002226 for Orsay virus, respectively) (Supplementary Fig. 11). These results suggest that high levels of variation in hyper-divergent regions likely enable diverse pathogen responses across the species. This hypothesis is supported by previous genetic mapping experiments in wild isolates that found substantial phenotypic variation in responses to the natural *C. elegans* pathogens *N. parisii*[30] and the Orsay virus[31], as well as responses to pathogens not found associated with *C. elegans* in nature[32].

In addition to pathogenic microbes in its natural habitat, *C. elegans* is often associated with diverse non-pathogenic microbes[33]. Because we do not know how natural genetic variation modulates responses to these microbes, we quantified the physiological responses of 92 *C. elegans* isotypes exposed to three naturally associated bacteria (JUb71: *Enterobacter sp.*, JUb85: *Pseudomonas putida*, and JUb87: *Buttiauxella agrestis*), which do not affect the growth rate of the N2 strain or induce stress response genes[33], and performed genome-wide association mappings (Methods, Supplementary Data 4). We identified a number of genomic regions that are associated with responses to these microbes (Supplementary Fig. 12), highlighting the role of genetic variation on modulating responses to non-pathogenic microbes. Furthermore, we found that hyper-divergent regions are overrepresented in genomic regions associated with these microbial responses (hypergeometric test; *p*-value = $2.6e^{-38}$). Taken together, these results establish a link between hyper-divergent regions and physiological responses to naturally associated microbes, strongly suggesting these regions are broadly involved in modulating how *C. elegans* senses and responds to its environment.

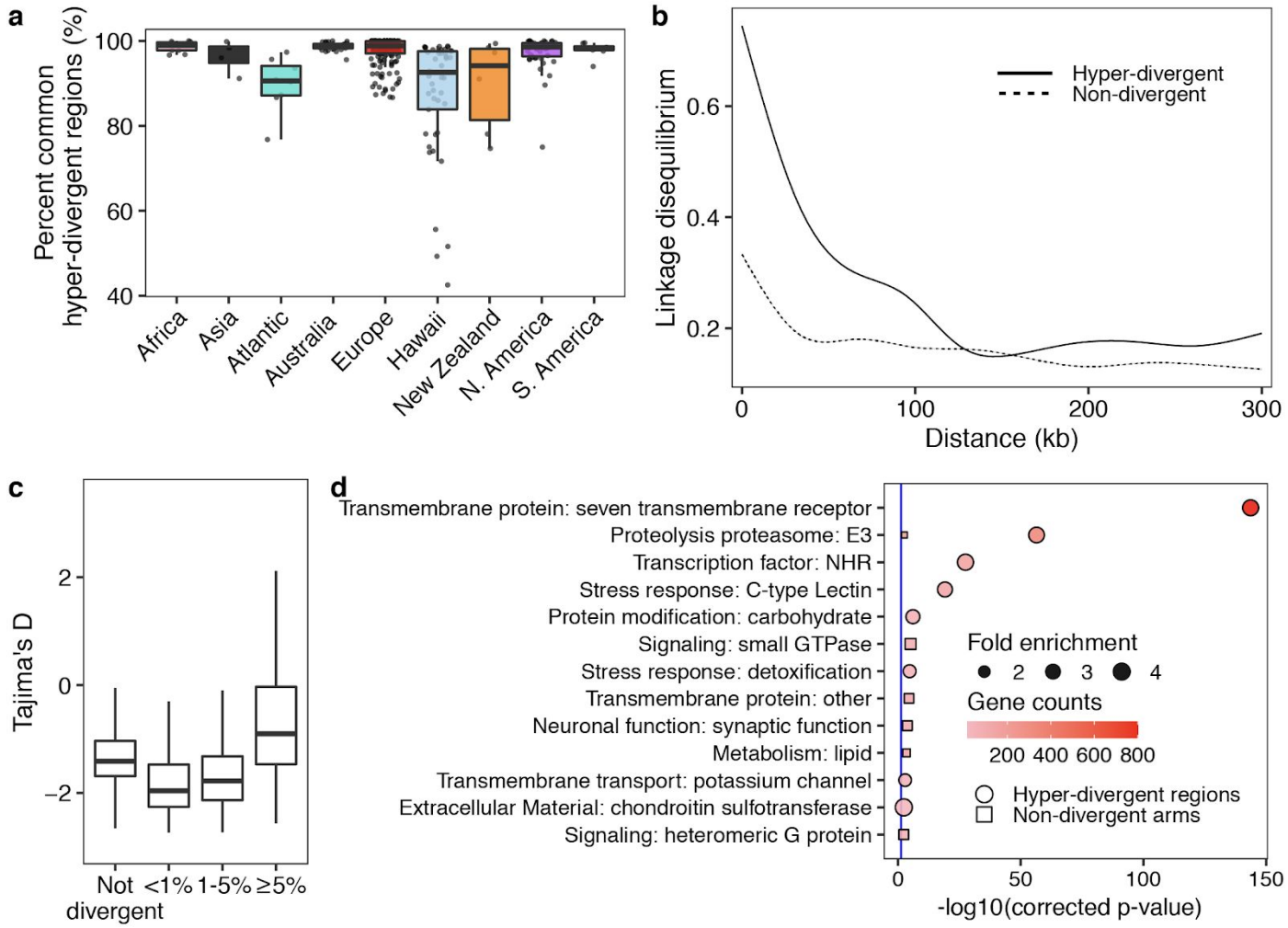



**Fig. 3 | Balancing selection has maintained hyper-divergent haplotypes enriched in environmental-response genes**

(a) Tukey box plots of the total amount of genome covered by common hyper-divergent regions are shown with data points plotted behind. Wild isotypes are grouped by their geographic origin. Each point corresponds to one of the 328 isotypes, and genome-wide fractions of common hyper-divergent regions are shown on the y-axis. The horizontal line in the middle of the box is the median, and the box denotes the 25th to 75th quantiles of the data. The vertical line represents the 1.5x interquartile range.

(b) Linkage disequilibrium (LD) decay in hyper-divergent regions and the rest of the genome is shown. The solid line and the dotted line are smoothed lines (generalized additive model (GAM) fitting) for LD decay in hyper-divergent regions in autosomal arms and non-divergent regions in autosomal arms, respectively. LD ($r^2$) between at least ten pairs of single nucleotide variants (SNVs) located in species-wide hyper-divergent regions in autosomal arms (n = 282,590) and LD between at least ten pairs of SNVs located in non-divergent regions in autosomal arms (n = 297,864) were used to generate the model. Note that 99.9% confidence intervals are represented around each smoothed line as grey bands, but not visible because of the narrow range of the intervals. Distances between SNVs are shown on the x-axis, and LD statistics ($r^2$) are shown on the y-axis.

(c) Tukey box plots of the Tajima's D statistics for 1 kb genomic bins in non-divergent regions and hyper-divergent regions. Genomic bins are grouped by the percent of 327 non-reference wild isotypes that are classified as divergent (rare < 1%, 1% ≤ intermediate < 5%, common ≥ 5%). Tajima's D statistics are plotted on the y-axis. The horizontal line in the middle of the box is the median, and the box denotes the 25th to 75th quantiles of the data. The vertical line represents the 1.5x interquartile range.

(d) Gene-set enrichment for autosomal arm regions (square) and hyper-divergent regions (circle) are shown. Nine annotations in WormCat category 2 (Methods) that are significantly enriched in control regions

or hyper-divergent regions (circle) or both are shown on the y-axis. Bonferroni-corrected significance values for gene-set enrichment are shown on the x-axis. Sizes of squares and circles correspond to the fold enrichment of the annotation, and colors of square and circle correspond to the gene counts of the annotation. The blue line shows the Bonferroni-corrected significance threshold (corrected p-value = 0.05).

**Hyper-divergent haplotypes contain potentially ancient genetic diversity**

A common feature of the hyper-divergent regions is that short-read sequencing coverage is lower than the average genome-wide coverage (47% less coverage, on average), suggesting that divergence in these regions might be high enough to prevent short reads from accurately aligning to the reference genome. Therefore, we took advantage of the 15 high-quality genomes we assembled to assess the content of these hyper-divergent regions. Strikingly, we find that these regions contain multiple hyper-divergent haplotypes that contain unique sets of genes and alleles that have substantially diverged at the amino acid level. For example, a hyper-divergent region on chromosome II (II:3,667,179-3,701,405 in the N2 reference genome) contains three distinct hyper-divergent haplotypes across the 16 isotypes for which we have high-quality assemblies (Fig. 4a). In this region, the reference isotype (N2) shares a haplotype with three wild isotypes and contains 11 protein-coding genes, including six GPCRs (*srx-97*, *srx-98*, *srx-101*, *srx-102*, *srx-104*, and *srx-105*). A second haplotype is shared among 11 wild isotypes and contains 20 protein-coding genes, including ten that are not conserved in the reference haplotype. The third haplotype is found only in a single isotype (DL238) and contains 17 protein-coding genes, including seven that are not present in the reference haplotype. For the genes that are conserved across all three haplotypes, alleles in different hyper-divergent haplotypes commonly show less than 95% amino acid identity (*e.g. F19B10.10* has an average between-haplotype identity of 88.3%, while *srx-97*, which lies outside the divergent region, has an average between-haplotype identity of 98.4%; Fig. 4b). The relationships inferred for this region using short-read SNP data were consistent with those inferred using the long-read assemblies (Fig. 4a), allowing us to infer the haplotype composition of all non-reference isotypes (Supplementary Fig. 13a, Methods). We find that a total of 59 isotypes contain the reference haplotype, 267 isotypes contain the second divergent haplotype, and one other isotype shares the DL238 haplotype. Consistent with our hypothesis that these haplotypes have been maintained by long-term balancing selection, the phylogenetic relationships in this region do not reflect the species-wide relationships; the two high-frequency haplotypes are present in all three genetically distinct *C. elegans*

groups defined previously (Fig. 4c, Supplementary Fig. 13a). In other regions of the genome, we find different numbers of hyper-divergent haplotypes across the 16 isotypes. For example, at the *peel-1 zeel-1* incompatibility locus[34] on chromosome I (I:2,318,291-2,381,851 in the N2 reference genome), we find two distinct hyper-divergent haplotypes across the 16 isotypes (Extended Data Fig. 5, Supplementary Fig. 13b). By contrast, we find seven distinct hyper-divergent haplotypes across the 16 isotypes, each with their own unique complement of loci (Extended Data Fig. 6, Supplementary Fig. 13c), in a region on the right arm of chromosome V (V:20,193,463-20,267,244 in the N2 reference genome) that contains three F-box loci (*fbxa-114*, *fbxa-113*, and *fbxb-59*).

To contextualize the high divergence we observe in hyper-divergent regions, we compared the level of divergence between hyper-divergent haplotypes to that observed between closely related *Caenorhabditis* species. As *C. elegans* lacks a closely related sister species (Fig. 4d), we compared the amino acid identities of *C. elegans* hyper-divergent alleles with the divergence between their orthologs in *Caenorhabditis briggsae* and *Caenorhabditis nigoni*, a closely related pair of species that are believed to have diverged from each other approximately 3.5 million years ago[35]. Notably, the average identity between hyper-divergent alleles within *C. elegans* is comparable to the divergence of their orthologs between *C. briggsae* and *C. nigoni* (mean identity in divergent regions of 97.7% and 96.4%, respectively, compared with a mean identity in non-divergent regions of 99.6% and 97.0%, respectively; Fig. 4e). Taken together, these results suggest that diversity in these regions might have been maintained for millions of years and could possibly predate the evolution of selfing in *C. elegans*, which is estimated to have occurred within the last four million years[36].

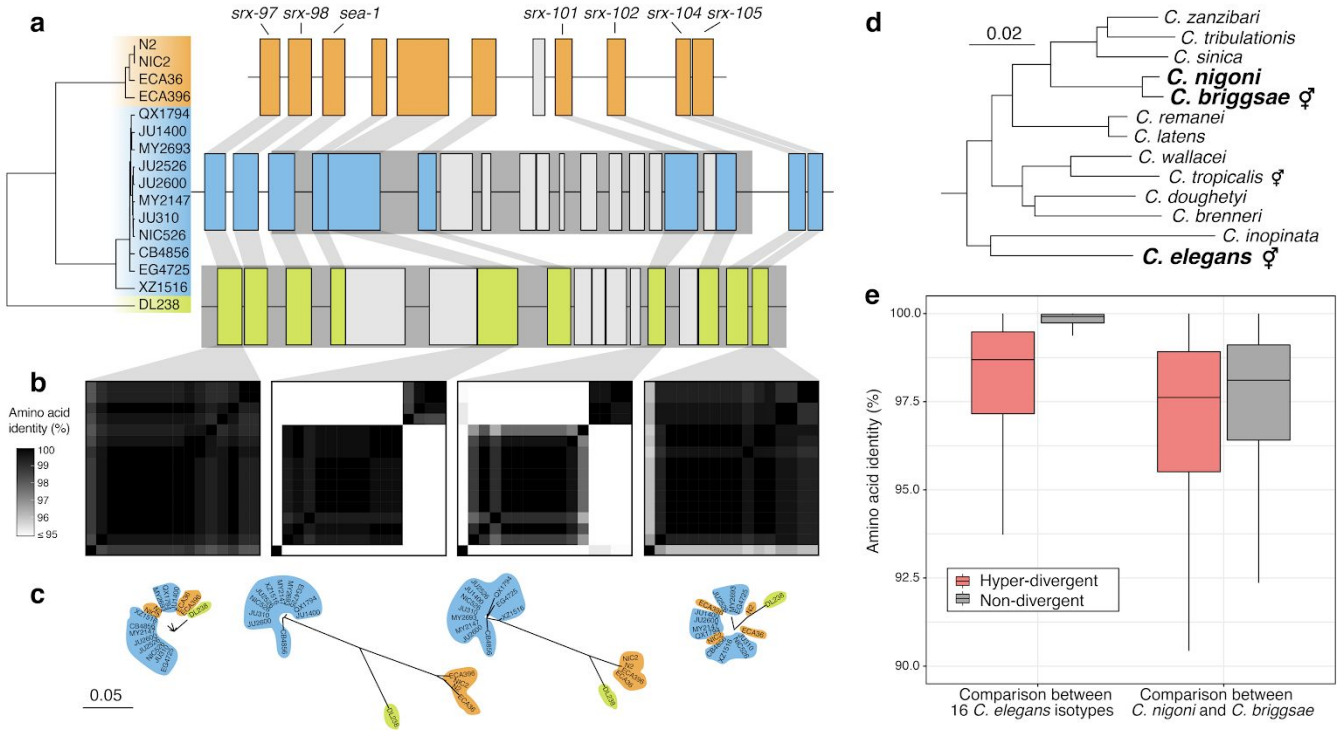



**Fig. 4 | Hyper-divergent haplotypes contain ancient genetic diversity**
(a) The protein-coding gene contents of three hyper-divergent haplotypes in a region on the left arm of chromosome II (II:3,667,179-3,701,405 of the N2 reference genome). The tree was inferred using SNVs and coloured by the inferred haplotypes. For each distinct haplotype, we chose a single isotype as a haplotype representative (orange haplotype: N2, blue haplotype: CB4856, green haplotype: DL238) and predicted protein-coding genes using both protein-based alignments and *ab initio* approaches (see Methods). Protein-coding genes are shown as boxes; those genes that are conserved in all haplotypes are coloured based on their haplotype, and those genes that are not are coloured light gray. Dark grey boxes behind loci indicate the coordinates of the hyper-divergent regions. Genes with locus names in N2 are highlighted.
(b) Heatmaps showing amino acid identity for alleles of four loci (*srx-97*, *F19B10.10*, *srx-101*, and *srx-105*). Percentage identity was calculated using alignments of proteins sequences from all 16 isotypes. Heatmaps are ordered by the SNV tree shown in (a).
(c) Maximum-likelihood gene trees of four loci (*srx-97*, *F19B10.10*, *srx-101*, and *srx-105*) inferred using amino acid alignments. Trees are plotted on the same scale (scale shown; scale is in amino acid substitutions per site). Isotype names are coloured by their haplotype.
(d) *Caenorhabditis* phylogeny showing relationships within the *Elegans* subgroup[37]. The positions of *C. elegans*, *C. briggsae*, and *C. nigoni* are highlighted. Species that reproduce via self-fertilisation are indicated. Scale is in amino acid substitutions per site.
(e) Tukey box plots showing amino acid identity of 8,741 genes that are single-copy and present in all 16 *C. elegans* isotypes, *C. briggsae*, and *C. nigoni*. Identities of alleles and their orthologs are shown separately for hyper-divergent and non-divergent regions.

## Hyper-divergent regions are common features in the genomes of selfing *Caenorhabditis* species

Selfing has evolved at least three times independently in the *Caenorhabditis* genus and is the main reproductive mode for *C. elegans*, *C. briggsae*, and *C. tropicalis*[38,39] (Fig. 4d). We hypothesized that long-term balancing selection might have maintained punctuated hyper-divergent regions in other selfing species. We used our short-read classification approach to identify hyper-divergent regions in the genomes

of 35 wild *C. briggsae* strains[35]. In agreement with our hypothesis, we find that hyper-divergent regions are widespread in the genomes of wild *C. briggsae* strains (Extended Data Fig. 7, Methods). Moreover, we find that the same regions are divergent in strains from the 'Tropical' *C. briggsae* clade and divergent strains from other clades[35]. Therefore, it is likely that the same evolutionary processes that have maintained ancient genetic diversity in *C. elegans* have also shaped the genome of *C. briggsae* and that hyper-divergent regions are common features in the genomes of selfing *Caenorhabditis* species.

## Discussion

Theory and empirical evidence show that predominantly selfing species have less genetic diversity than obligately outcrossing species[16]. It follows that reduced levels of genetic diversity in selfing species would limit the adaptive potential and range of ecological niches that the species can inhabit[4]. However, *C. elegans* and other selfing *Caenorhabditis* species are globally distributed, found in diverse habitats, and often share niches with outcrossing nematode species[38,40]. We provide evidence that ancient genetic diversity in hyper-divergent regions likely contributes to the distribution of *C. elegans* across diverse niches. We found that these genomic regions are overrepresented in quantitative trait loci correlated with responses to microbes that are naturally associated with *C. elegans* and significantly enriched for genes that modulate responses to bacterial foods, competitors, phoretic carriers, predators, and pathogens in wild habitats[41], including GPCRs, C-type lectins, and the *pals* gene family. Two chemosensory genes we identified as hyper-divergent, *srx-43* and *srx-44*, encode for GPCRs that were previously shown to contain two ancient haplotypes maintained by long-term balancing selection, likely to enable distinct density-dependent foraging strategies among *C. elegans* isotypes[19,42]. The hyper-divergent regions we present here also include other loci previously characterized to be under long-term balancing selection and have profound impacts on *C. elegans* physiology and ecology, including *glc-1*[43], *peel-1 zeel-1*[34], and *sup-35 pha-1*[44]. However, these loci only account for four of the 366 distinct hyper-divergent regions we identified, suggesting that we still have much to learn about the role of these regions on *C. elegans* ecology and physiology.

We hypothesize that hyper-divergent haplotypes represent ancestral genetic diversity that has been maintained by ancient balancing selection since the evolution of selfing in *C. elegans*, which is believed to

have occured in the last four million years[36]. The amino acid divergence among *C. elegans* hyper-divergent haplotypes is similar to the divergence between *C. briggsae* and *C. nigoni*, two species that diverged approximately 3.5 million years ago[35], suggesting that these regions have been maintained for millions of years. Moreover, outcrossing *Caenorhabditis* species typically have extremely high levels of genetic diversity[45,46]. Assuming the outcrossing ancestor of *C. elegans* was similarly diverse, these hyper-divergent regions might represent the last remnants of ancestral genetic diversity that has otherwise been lost because of the long-term effects of selfing. Similar observations have been reported in the *Capsella* genus, where the selfing species *Capsella rubella* shares variants with its outcrossing sister species *Capsella grandiflora*[47]. These trans-specific polymorphisms are predominantly found at loci involved in immune response[20], which match our findings in *C. elegans*. Although it is likely that hyper-divergent regions also exist in outcrossing *Caenorhabditis* species, identifying them in selfers is expected to be easier for two reasons. First, selfing leads to a reduction in the effective recombination rate, increasing the overall footprint of balancing selection[48,49]. Second, long-term selfing leads to an overall reduction of genome-wide diversity, making the high polymorphism at balanced regions of the genome more significant[5]. These processes likely explain the punctuated, hyper-divergent nature of the regions we found in *C. elegans*.

An alternative explanation for the origin of these regions is adaptive introgression from divergent taxa, which was recently shown to explain the existence of large, divergent haplotypes that underlie ecotypic adaptation in sunflowers[50], and characterized in a wide-range of other species[51]. Although possible, the presence of up to seven *C. elegans* hyper-divergent haplotypes at a single locus suggests that recent introgression is unlikely to be an exclusive source of hyper-divergent haplotypes. Because a closely related *C. elegans* sister species has not been identified yet, it is not possible to distinguish between retained ancestral polymorphism and recent introgression. However, as similar population-wide variant datasets become available for *C. briggsae*, it will be possible to test whether hyper-divergent haplotypes are shared with the closely related outcrossing species *C. nigoni* and if the patterns of divergence are consistent with introgression, retained ancestral genetic diversity, or both.

Regardless of their origin, the existence of these regions in *C. elegans* has important implications for how we understand the genetic and genomic consequences of selfing. It has been proposed that the evolution of selfing represents an evolutionary “dead-end”, whereby the reduction in genetic diversity, and

therefore adaptive potential, of a species will eventually lead to extinction[52]. However, our findings suggest that it is possible to avoid this fate by maintaining a substantial fraction of the ancestral genetic diversity at key regions of the genome.

## Methods

### Strains

Nematodes were reared at 20ºC using *Escherichia coli* bacteria (strain OP50) grown on modified nematode growth medium (NGMA)[53], containing 1% agar and 0.7% agarose to prevent animals from burrowing. All 609 wild *C. elegans* strains are available on CeNDR[54] and accompanying metadata (Supplementary Table 1).

### Sequencing and isotype characterization

*Sequencing*. To extract DNA, we transferred nematodes from two 10 cm NGMA plates spotted with OP50 into a 15 ml conical tube by washing with 10 mL of M9. We then used gravity to settle animals in a conical tube, removed the supernatant, and added 10 mL of fresh M9. We repeated this wash method three times over the course of one hour to serially dilute the *E. coli* in the M9 and allow the animals time to purge ingested *E. coli*. Genomic DNA was isolated from 100 to 300 µl nematode pellets using the Blood and Tissue DNA isolation kit (cat#69506, QIAGEN, Valencia, CA) following established protocols[55]. The DNA concentration was determined for each sample with the Qubit dsDNA Broad Range Assay Kit (cat#Q32850, Invitrogen, Carlsbad, CA). The DNA samples were then submitted to the Duke Sequencing and Genomic Technologies Shared Resource per their requirements. The Illumina library construction and sequencing were performed at Duke University using KAPA Hyper Prep kits (Kapa Biosystems, Wilmington, MA) and the Illumina NovaSeq 6000 platform (paired-end 150 bp reads). The raw sequencing reads for strains used in this project are available from the NCBI Sequence Read Archive (Project PRJNA549503).

*Isotype characterization*. Raw sequencing reads from 609 wild strains were trimmed using the Trimmomatic (v0.36)[56] to remove low-quality bases and adapter sequences. Following trimming, we called SNVs using the BCFtools (v.1.9)[57] and the following filters: Depth (DP) ≥ 3; Mapping Quality (MQ) > 40; Variant quality (QUAL) > 30; Allelic Depth (FORMAT/AD)/Num of high-quality bases (FORMAT/DP) ratio > 0.5. We classified two or more strains as the same isotype if they have the same call at 99.9% of all sites called across the full panel of wild strains. If a strain did not meet this criterion, we considered it as a unique isotype. Newly assigned isotypes were added to CeNDR[54].

### Sequence alignments and variant calling

After isotypes are assigned, we used alignment-nf with BWA (v0.7.17-r1188)[58,59] to align trimmed sequence data for distinct isotypes to the N2 reference genome (WS245)[60]. Variants were called using GATK4 (v4.1.0)[61]. First, gVCFs were generated for each isotype using the *HaplotypeCaller* function in GATK4 with the following parameters: *--max-genotype-count*=3000 and *--max-alternate-alleles*=100, using the WS245 N2 reference genome. Next, individual isotype gVCFs were merged using the *MegeVcfs* function in GATK4 and imported to a genomics database using the *GenomicsDBImport* function. Genotyping of the gVCFs was performed using the *GenotypeGVCFs* function in GATK4 with the following parameter *--use-new-qual-calculator*. The 328 isotype cohort VCF was annotated using SnpEff[62] and an annotation database that was built with the WS261 gene annotations[62]. Following VCF annotation, we applied soft filters (QD < 5.0, SOR > 5.0, QUAL < 30.0, ReadPosRankSum < -5.0, FS > 50.0, DP < 5) to the VCF variants using the *VariantFiltration* function in GATK4. We applied a final isotype-specific soft filter called *dv_dp* using the bcftools *filter* function, which required the alternate allele depth to at least 50% of the total read depth for an individual isotype. All variant sites that failed to meet the variant-level filter criteria were removed from the soft-filtered VCF, and all isotype-level variants that did not meet the *dv_dp* criteria were set to missing. Finally, we removed sites that had more than 5% missing genotype data or more than 10% of samples were called heterozygous.

### Genetic relatedness

*Similarity analysis*. Using BCFtools (v.1.9) with the command filter -i N_MISSING=0, we filtered the high-quality VCF file and generated a VCF file (complete-site VCF) containing 963,027 biallelic SNVs that are genotyped for all 328 *C. elegans* wild isotypes. We used the vcf2phylip.py script[63] to convert the complete-site VCF file to the PHYLIP format. The distance matrix and unrooted neighbor-joining tree were made from this PHYLIP file using *dist.ml* and *NJ* function using the phangorn (v2.5.5) R package[64]. The tree was visualized using the ggtree (version 1.16.6) R package[65].

*Principal component analysis.* The *smartpca* executable from the EIGENSOFT (v6.1.4)[11,66] was used to perform principal component analysis. We performed analysis with the complete-site VCF with or without removing outlier isotypes to analyze the population structure with highly genetically divergent isotypes and many related swept isotypes. When analyzing the population without removing outlier isotypes, we used the following parameters: altnormstyle: NO, numoutevec: 50, familynames: NO. When analyzing the population with outlier isotype removal, we set numoutlieriter to 15.

**Population genomic analyses**

*Haplotype analysis*. We determined identity-by-descent (IBD) of genome segments using IBDSeq (version r1206)[67] run on the complete-site VCF with the following parameters:minalleles = 0.01, r2window = 1500, ibdtrim = 0, r2max = 0.3 for genome-wide haplotype analysis and minalleles = 0.01, r2window = 1500, ibdtrim = 0, r2max = 1 for hyper-divergent regions. IBD segments were then used to infer haplotype structure among isotypes as described previously[13]. To define the swept haplotype, we first identified the most common haplotype found on each chromosome that passed following per chromosome filters: total length >1 Mb; total length/maximum population-wide haplotype length > 0.03. Second, we calculated the average fractions of each chromosome covered by the most common haplotype across 328 wild isotypes. We defined the most common haplotype in chromosome I, IV, V, and X, of which the species-wide average fraction of the most common haplotype is greater than 30%, as the swept haplotype for each chromosome.

*Population genetics*. We only considered bi-allelic SNVs to calculate population genomic statistics. Tajima's *D*, Watterson's theta, and Pi were all calculated using scikit-allel[68]. Each of these statistics was calculated for the same non-overlapping 1000 bp windows as hyper-divergent regions (described below).

*Linkage disequilibrium (LD) decay*. We filtered the complete-site VCF file using BCFtools (v1.9) with the command view -q 0.05:minor and generated a VCF file (MAF-05 VCF) containing 123,830 SNVs of which minor allele frequencies are greater than or equal to 5%. Then, we selected 41,368 SNVs on autosomal arms (MAF-05-autoarm VCF) and split the MAF-05-autoarm VCF by the location of SNVs; MAF-05-autoarm-div VCF with 17,419 SNVs within hyper-divergent autosomal arm and MAF-05-autoarm-nondiv VCF with 23,949 SNVs within the non-divergent autosomal arms. We analyzed LD decay by running PopLDdecay (v3.31)[69] on both MAF-05-autoarm-div VCF and MAF-05-autoarm-nondiv VCF with the default parameters (MaxDist=300, Het=0.88, Miss=0.25).

**Characterization of hyper-divergent regions**

*C. elegans*. To characterize hyper-divergent regions across 327 non-reference wild *C. elegans* isotypes, we used the sequencing depth and variant (SNV/indel) counts information from short-read alignments for each isotype. We performed a sliding window analysis with a 1 kb window size and a 1 kb step size for each of 327 wild isotype genomes. First, we counted the number of SNV and indels (variant counts) for each window from the high-quality VCF using BEDtools (v2.27.1)[70] with the command coverage -counts. Second, we analyzed the average sequencing depth of each window using mosdepth (v0.2.3)[71]; then we calculated the relative sequencing depth of each window to the genome-wide average depth (coverage fraction = average sequencing depth of the window/genome-wide average depth). We classified each window as hyper-divergent if its variant counts ≥ 16 or coverage fraction < 35% or both; we also classified windows that are flanked by hyper-divergent windows as hyper-divergent. Third, we clustered contiguous hyper-divergent windows and defined clusters that are greater than or equal to 9 kb of N2 reference genome length as hyper-divergent regions[18]. Additionally, we joined these clusters if the distance between two clusters is less or equal than 5 kb. The thresholds for variant counts and coverage fraction were chosen based on alignments of long-read assemblies of 15 wild isotypes (Supplementary Fig. 5). In summary, we reduced both false classifications of hyper-divergent regions and non-divergent regions by comparing coverage and identity of alignments of long-read assemblies. At the same time, we found that too strict thresholds (high variant counts threshold and low coverage fraction) increase false classification of hyper-divergent regions into non-divergent regions and underestimate the total size of hyper-divergent regions (2.8 Mb) identified in CB4856 previously[18]. With selected threshold parameters (variant counts ≥ 16 and coverage fraction < 35%), we characterized a similar size of hyper-divergent regions (3.2 Mb) in CB4856. Additionally, we confirmed that selected parameters do not detect any hyper-divergent region from short-read alignments of N2 reference strain to its own reference genome. Characterization of hyper-divergent regions is summarized in Extended Data Fig. 4b. To classify species-wide hyper-divergent

regions, we identified 1 kb genomic bins that were classified as divergent in at least one isotype and grouped contiguous bins as a hyper-divergent region. To compare small variant (SNVs/indels) density between divergent and non-divergent regions for chromosomal centers, arms, and tips (Supplementary Table 5), we used previously defined genomic coordinates for centers, arms, and tips of six chromosomes[14]. For bins with no isotype classified as hyper-divergent, we measured the variant density as the average variant density of all 328 wild isotypes. For bins with at least one isotype classified as hyper-divergent, we measured the variant density as the average variant density of wild isotypes that are classified as hyper-divergent for each bin. We calculated the percent of 327 non-reference wild isotypes that is classified as hyper-divergent (percent divergence) for each 1 kb bin, and classified each bin into one of three frequency groups based on its percent divergence: rare < 1%, 1% ≤ intermediate < 5%, common ≥ 5%.

*C. briggsae*. We performed a sliding window analysis with a 1 kb window size and a 1 kb step size for each of 36 non-reference wild *C. briggsae* genomes[35]. We only used variant counts to classify hyper-divergent regions. Because the *C. briggsae* genome-wide variant density is 1.55 greater than *C. elegans*, we used a modified variant count threshold (variant counts ≥ 24). We also classified windows that were flanked by hyper-divergent windows as hyper-divergent.

**Gene-set enrichment analysis**

We analyzed the gene-set enrichment of hyper-divergent regions using web-based WormCat[72] that contains a near-complete annotation of genes from the *C. elegans* reference strain N2. We also performed a conventional gene ontology (GO) enrichment analysis using the clusterProfiler (v3.12.0) R package[73] and org.Ce.eg.db: Genome-wide annotation for Worm[74]. Because the majority of hyper-divergent regions are found on autosomal arms (Fig. 2a, Supplementary Table 5), we used gene-set enrichment on the autosomal arms as a control dataset[14].

**Preparation of natural bacteria**

The following natural bacteria were isolated by Marie-Anne Felix from nematode samples, JUb71 (*Enterobacter sp.*), JUb85 (*Psuedomonas putida*), or JUb87 (*Buttiauxella agrestis*). Each natural bacterium was grown overnight, diluted 1:500 in a 30 L culture of LB. Cultures were grown at 30°C until late log phase ($OD_{600}$ = 0.8). The full culture was pelleted by centrifugation, washed twice with distilled water, pelleted again, resuspended in distilled water at $OD_{600}$=100, and then aliquoted for freezing at -80°C.

**High-throughput measurement of growth in natural foods**

Animals were fed either HB101 lysate[75] as a control or the natural bacteria at an $OD_{600}$=15 and subjected to a previously developed high-throughput fitness assay (HTA) [76]. Briefly, strains are passaged for four generations and then bleach-synchronized and aliquoted to 96-well microtiter plates at approximately one embryo per microliter in K medium [77]. Embryos were then hatched overnight to the L1 larval stage. The following day, hatched L1 animals are fed HB101 bacterial lysate (Pennsylvania State University Shared Fermentation Facility, State College, PA) at a final concentration of 5 mg/ml and grown to the L4 stage after two days at 20°C. Three L4 larvae are then sorted using a large-particle flow cytometer (COPAS BIOSORT, Union Biometrica, Holliston, MA) into microtiter plates that contain HB101 lysate at 10 mg/ml, K medium, 31.25 µM kanamycin, or natural bacteria. The animals are then grown for four days at 20°C. Prior to the measurement of fitness parameters from the population, animals were treated with sodium azide (50 mM) to straighten their bodies for more accurate length measurements. Traits that are measured by the BIOSORT include brood size, animal length (time of flight or TOF), optical density (extinction or EXT), and fluorescence.

Phenotype data generated using the BIOSORT were processed using the R package easysorter, which was specifically developed for processing this type of data set [78,79]. Briefly, the function *read_data*, reads in raw phenotype data, runs a support vector machine to identify and eliminate bubbles. Next, the *remove_contamination* function eliminates any wells that were contaminated prior to scoring population parameters for further analysis. Contamination is assessed by visual inspection. The sumplate function is then used to generate summary statistics of the measured parameters for each animal in each well. These summary statistics include the 10th, 25th, 50th, 75th, and 90th quantiles for TOF. Measured brood sizes are normalized by the number of animals that were originally sorted into the well. After summary statistics

for each well are calculated, the regress(assay = TRUE) function in the easysorter package is used to fit a linear model with the formula (phenotype ~ assay) to account for any differences between assays. Next, outliers are eliminated using the *bamf_prune* function. This function eliminates strain values that are greater than two times the IQR plus the 75th quantile or two times the IQR minus the 25th quantile, unless at least 5% of the strains lie outside this range. Finally, bacteria-specific effects are calculated using the regress(assay = FALSE) function from easysorter, which fits a linear model with the formula (phenotype ~ HB101 phenotype) to account for any differences in population parameters present in control HB101 conditions.

**Genome-wide association mappings and enrichment analysis**

Genome-wide association (GWA) mapping was performed using phenotype data from 92 *C. elegans* isotypes (Supplementary Data 9). Genotype data were acquired from the latest imputed VCF release (Release 20180527) from CeNDR that was imputed as described previously[55]. We used BCFtools[57] to filter variants that had any missing genotype calls and variants that were below 5% minor allele frequency in the phenotyped population. We used PLINK v1.9 (Purcell et al., 2007; Chang et al., 2015) to LD-prune the genotypes at a threshold of r2 <0.8, using *--indep-pairwise* 50 10 0.8. This genotype set consisted of 64,053 markers that were used to generate the realized additive kinship matrix using the *A.mat* function in the *rrBLUP* R package[80]. These markers were also used for genome-wide association mapping. However, because these markers still have substantial LD within this genotype set, we performed eigen decomposition of the correlation matrix of the genotype matrix using *eigs_sym* function in Rspectra package[81]. The correlation matrix was generated using the cor function in the correlateR R package[82]. We set any eigenvalue greater than one from this analysis to one and summed all of the eigenvalues. This number was 537, which corresponds to the number of independent tests within the genotype matrix and was used to determine the significance threshold for significant QTL[83]. We used the GWAS function in the rrBLUP package to perform genome-wide mapping with the following command: rrBLUP::GWAS (pheno = trait_file, geno = Pruned_Markers, K = KINSHIP, min.MAF = 0.05, n.core = 1, P3D = FALSE, plot = FALSE). Genomic regions associated with the bacteria responses were determined as previously defined[84], but with +/- 150 SNVs from the rightmost and leftmost markers above the Bonferroni significance threshold. The workflow for performing GWA mapping can be found on github.com/elifesciences-publications/cegwas2-nf.

To determine if the detected QTL were enriched for hyper-divergent regions, we only considered hyper-divergent genomic regions present in the 92 phenotyped strains. Next, we binned each of the hyper-divergent regions into the same 1 kb bins that were used to define them and only considered 1 kb bins that were classified as hyper-divergent in at least 5% of the strains, which corresponds to the minor allele frequency threshold we used for GWA mapping. We performed the same 1 kb binning procedure for each unique QTL we detected. Next, we performed an intersection between the QTL and hyper-divergent 1 kb bins using the *bed_intersect* function in the valr R package[85] to calculate the number of 1 kb QTL bins that overlapped with hyper-divergent 1 kb bins. We also performed the inverse operation with the *bed_intersect* (with *invert* = TRUE) function to identify the extent to which identified QTL overlapped with the non-hyper-divergent genome. Finally, we performed a hypergeometric test using the *phyper* function in R[86] with the following parameters: *x* = the number of 1 kb QTL bins that overlapped with hyper-divergent 1 kb bins, *m* = the total number of 1 kb hyper-divergent bins in the genome, *n* = the total number of 1 kb bins that are not classified as hyper-divergent, *k* = the total number of 1 kb QTL bins that were detected, and *lower.tail* = *FALSE*.

**Pacific Biosciences continuous long-read sequencing**

To extract DNA, we transferred nematodes from twelve 10 cm NGMA plates spotted with OP50 into a 50 ml conical tube by washing with 30 mL of M9. We then used gravity to settle animals for 1.5 hours, removed the supernatant, and added 15 mL of fresh M9. We allowed nematodes to settle for 1 hour, removed the supernatant, and transferred nematodes to a microfuge tube using a Pasteur pettle. We settled animals for an additional 1 hour and any supernatant was removed. We stored pellets at -80°C prior to DNA extraction. Genomic DNA was isolated from 400 to 500 µl nematode pellets using the MagAttract HMW DNA kit (cat#67563, QIAGEN, Valencia, CA). The DNA concentration was determined for each sample with the Qubit dsDNA Broad Range Assay Kit (cat#Q32850, Invitrogen, Carlsbad, CA). The DNA samples were

then submitted to the Duke CSequencing and Genomic Technologies Shared Resource per their requirements. The Pacific Biosciences library construction and sequencing were performed at Duke University using the Pacific Biosciences Sequel platform. Quality control was performed with the Qubit dsDNA Broad Range Assay Kit and Agilent Tapestation. Some samples had 30-50 kb fragment sizes, and a 15 kb cutoff was used for size selection of these samples during library prep. A 20 kb cutoff was used for the other samples. The raw sequencing reads for strains used in this project are available from the NCBI Sequence Read Archive (Project PRJNA647911).

**Long-read genome assembly**

We extracted PacBio reads in FASTQ format from the subread BAM files using the PacBio bam2fastx tool (version 1.3.0; available at https://github.com/PacificBiosciences/bam2fastx). For each of the 14 wild isotypes, we assembled the PacBio reads using three genome assemblers: wtdbg2 (version 0.0; using the parameters *-x sq -g 102m*)[87], flye (version 2.7; using the parameters *--pacbio-raw -g 102m*)[88], and Canu (version 1.9; using the parameters genomeSize=102m -pacbio-raw)[89]. For each assembly, we assessed the biological completeness using BUSCO[90] (version 4.0.6; using the options *-l nematoda_odb10 -m genome*) and contiguity by calculating a range of numerical metrics using scaffold_stats.pl (available here: https://github.com/sujaikumar/assemblage)[90]. We selected the Canu assemblies as our final assemblies because they had both high contiguity and high biological completeness. To correct sequencing errors that remained in the Canu assemblies, we aligned short-read Illumina data for each wild isotype to the corresponding assembly using BWA-MEM[59] (version 0.7.17) and provided the BAM files to Pilon for error correction (version 1.23; using the parameters *--changes --fix bases*)[91]. To remove assembled contigs that originated from non-target organisms (such as the *E. coli* food source), we screened each assembly using taxon-annotated GC-coverage plots as implemented in BlobTools[92]. Briefly, we aligned the PacBio reads to the assembly using minimap2[93] (version 2.17; using the parameters *-a -x map-pb*) and sorted then indexed the BAM files using SAMtools[94] (version 1.9). We searched each assembled contig against the NCBI nucleotide database ('nt') using BLASTN[95] (version 2.9.0+; using the parameters *-max_target_seqs 1 -max_hsps 1 -evalue 1e-25*) and against UniProt Reference Proteomes[96] using Diamond (version 0.9.17; using the parameters *blastx --max-target-seqs 1 --sensitive --evalue 1e-25*)[97]. We removed contigs that were annotated as being of bacterial origin and that had coverage and percent GC that differed from the target genome. Genome assembly metrics are shown in Supplementary Table 3.

**Protein-coding gene prediction**

To generate *ab initio* gene predictions for all 14 assemblies and for a previously published long-read assembly for the Hawaiian isotype CB4856[17], we first used RepeatMasker[98] (ver 4.0.9; using the parameter *-xsmall*) to identify and soft-mask repetitive elements in each genome assembly using a library of known Rhabditida repeats from RepBase[99]. We predicted genes in the masked assemblies using AUGUSTUS (version 3.3.3; using the parameters *--genemodel=partial --gff3=on --species=caenorhabditis --codingseq=on*)[100]. We extracted protein sequences and coding sequences from the GFF3 file using the getAnnoFasta.pl script from AUGUSTUS. We assessed the biological completeness of each predicted gene set using BUSCO (version 4.0.6; using the options *-l nematoda_odb10 -m proteins*) using the longest isoforms of each gene only. Predicted protein-coding gene counts and BUSCO completeness scores are shown in Supplementary Table 3.

**Whole-genome alignments**

We aligned all 14 long-read assemblies generated here along with a long-read assembly for the Hawaiian isotype CB4856 to the N2 reference genome (WS255)[101] using NUCmer (version 3.1; using the parameters *--maxgap=500 --mincluster=100*)[102]. Coordinates and identities of the aligned sequences were extracted for the alignment files using the 'show-coords' function with NUCmer.

**Characterising gene contents of hyper-divergent haplotypes**

To characterise and compare the gene contents of hyper-divergent haplotypes, we used OrthoFinder [103] (version 2.3.11; using the parameter *-og*) to cluster the longest isoform of each protein-coding gene predicted from the genomes of 15 wild isolates and the N2 reference genome (WS270). We initially attempted to use the orthology assignments to identify alleles in the 15 wild isotypes for each region of

interest. However, gene prediction errors were pervasive, with many fused and split gene models, which caused incorrect orthology assignment and/or spurious alignments. To ensure the differences between hyper-divergent haplotypes were real biological differences and not gene prediction artefacts, we used the AUGUSTUS gene predictions and orthology assignments to identify coordinates of hyper-divergent haplotypes only. For a given region of interest, we identified genes in the reference genome that flanked the hyper-divergent regions and used the orthology assignments to identify the corresponding alleles in all 15 wild isolates. Using the coordinates of these genes, we extracted the sequences of intervening hyper-divergent haplotypes using BEDtools (version 2.29.2; using the parameters *getfasta -bed*)[70]. To identify genes that were conserved in the reference haplotype, we extracted the longest isoform for each protein-coding gene in this region from the N2 reference gene set and used exonerate[104] (version 2.4.0; using the parameters *--model p2g --showvulgar no --showalignment no --showquerygff no --showtargetgff yes*) to infer directly gene models for each of the 15 wild isotypes. To predict genes that were not present in the reference haplotype, we first used BEDtools (version v2.29.2; with the parameter *maskfasta*) to mask the coordinates of the exonerate-predicted genes with Ns. We then used AUGUSTUS (version 3.3.3; using the parameters *--genemodel=partial --gff3=on --species=caenorhabditis --codingseq=on*) to predict genes in the unmasked regions. To remove spurious gene predictions or transposable element loci, we searched the protein sequence of each AUGUSTUS-predicted gene against the N2 reference genome using BLASTP (version 2.9.0+; using the parameters *-max_target_seqs 1 -max_hsps 1*) and against the Pfam database[105] using InterProScan (version 5.35-74.0; using the parameters *-dp -t p --goterms -appl Pfam -f tsv*)[106]. Predicted genes that contained Pfam domains associated with transposable elements, or those genes that lacked sequence similarity to a known *C. elegans* protein sequence and a conserved protein domain were discarded. The coordinates of all curated predicted protein-coding genes were used to generate gene content plots (Fig 4a; Extended Data Fig. 5a; Extended Data Fig. 6a) using the ggplot R package[107]. To generate amino acid identity heatmaps and gene trees, we aligned the protein sequences of all conserved genes (those predicted with exonerate) using FSA (version 1.15.9)[108]. We inferred gene trees using IQ-TREE[109] (version 2.0.3), allowing the best-fitting substitution model to be automatically selected for each alignment[110]. Gene trees were visualised using the ggtree R package. We calculated percentage identity matrices from the protein alignments using a custom Python script (available at https://github.com/AndersenLab/Ce-328-pop-divergent). We then generated heatmaps, ordered by the strain relatedness tree inferred for the region, using the ggplot R package[107]. GFF files, protein alignments, gene trees, and percentage identity matrices for each characterised region are available at https://github.com/AndersenLab/Ce-328-pop-divergent. To construct dendrograms of hyper-divergent regions, we first converted the hard-filtered VCF to a gds object using the *snpgdsVCF2GDS* function in the SNPrelate package[111]. Next, we calculated the pairwise identity-by-descent fraction for all strains using the *snpgdsIBS* function in the *SNPrelate* R package, and performing hierarchical clustering on the matrix using the *snpgdsHCluster* function in the *SNPrelate* package. We visualized the strain relatedness using ggtree[65].

**Comparing divergence between *C. briggsae* and *C. nigoni***

To compare the divergence between hyper-divergent haplotypes with closely related *Caenorhabditis* species, we downloaded the protein sequences predicted from the genomes of *Caenorhabditis briggsae*[112] and *Caenorhabditis nigoni*[113] from WormBase (WS275). We clustered the longest isoform of each protein-coding gene for both species with the longest isoform of each protein-coding gene in all 16 *C. elegans* isotypes using OrthoFinder (version 2.3.11; using the parameter *-og*). We identified 8,741 orthogroups containing protein sequences that were present and single-copy in all 18 taxa and aligned their sequences using FSA (version 1.15.9). We calculated a percentage identity matrix for each protein alignment using a custom Python script (available at https://github.com/AndersenLab/Ce-328-pop-divergent). For each of the 15 wild isotypes, we partitioned the 8,741 genes into those that were classified as being divergent and those that were classified as being non-divergent by our short-read classification approach. We then extracted the percentage identity of the allele of interest and the N2 alleles, and also the percentage identity between the corresponding ortholog in *C. briggsae* and *C. nigoni*. For each gene, we calculated mean identity between all non-divergent alleles and the correspond N2 alleles (and between their orthologs in *C. briggsae* and *C. nigoni*) and the mean identity between all hyper-divergent alleles and the corresponding N2 allele (and between their orthologs in *C. briggsae* and *C. nigoni*).

## Data and code availability

All data sets and code for generating figures and tables are available on GitHub (https://github.com/AndersenLab/Ce-328-pop-divergent).

## Acknowledgments

We would like to thank members of the Andersen lab for providing comments on this manuscript. We especially thank Michael Ailion, Jean David, Robert Luallen, Nathalie Pujol, and citizen-scientists for their contributions of wild *C. elegans* strains to CeNDR. We also thank the Duke University School of Medicine for the use of the Sequencing and Genomic Technologies Shared Resource, which provided Pacific Biosystems long-read sequencing.

## Funding

This work was funded by an NSF CAREER award (1751035) and a Human Frontiers Science Program Award (RGP0001/2019). NIH grant ES029930 to E.C.A, M.V.R., and L.R.B. also funded this work. S.Z. received funding from The Cellular and Molecular Basis of Disease training program (T32GM008061) and the Rappaport Award for Research Excellence through the IBiS graduate program. A.K.W. is supported by the National Science Foundation Graduate Research Fellowship. Long-read sequencing of three isolates was funded by the National Institutes of Health R01 (GM117408) to L.R.B. and a T32 training grant for the University Program in Genetics and Genomics (GM007754). M.V.R. is supported by NIH grant GM121828. M.G.S. was supported by NWO domain Applied and Engineering Sciences VENI grant (17282).

## Author contributions

D.L., S.Z., and E.C.A. conceived and designed the study. D.L., S.Z., L.S., and E.C.A. analyzed the data and wrote the manuscript. Y.W., R.E.T., and D.E.C. performed whole-genome sequencing and isotype characterization for 609 wild *C. elegans* strains. R.E.T. performed long-read sequencing for 11 *C. elegans* wild isolates. R.C., A.K.W., and L.R.B. performed long-read sequencing for three *C. elegans* wild isolates. M.G.S., C.B., M.V.R., and M.-A.F. contributed wild isolates to the *C. elegans* strain collection and edited the manuscript. T.A.C. edited the manuscript.

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary information** is available for this paper.
**Correspondence and requests for materials** should be addressed to E.C.A.

## References


1. Heller, R. & Maynard Smith, J. Does Muller's ratchet work with selfing? *Genet. Res.* **32**, 289–293 (1978).
2. Charlesworth, D., Morgan, M. T. & Charlesworth, B. Mutation accumulation in finite outbreeding and inbreeding populations. *Genet. Res.* **61**, 39–56 (1993).
3. Charlesworth, B., Morgan, M. T. & Charlesworth, D. The effect of deleterious mutations on neutral molecular variation. *Genetics* **134**, 1289–1303 (1993).
4. Busch, J. W. & Delph, L. F. Evolution: Selfing Takes Species Down Stebbins's Blind Alley. *Curr. Biol.* **27**, R61–R63 (2017).
5. Charlesworth, D. & Wright, S. I. Breeding systems and genome evolution. *Curr. Opin. Genet. Dev.* **11**, 685–690 (2001).
6. Charlesworth, D. Effects of inbreeding on the genetic diversity of populations. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **358**, 1051–1070 (2003).
7. Barrett, S. C. H., Arunkumar, R. & Wright, S. I. The demography and population genomics of evolutionary transitions to self-fertilization in plants. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **369**, (2014).
8. Cutter, A. D., Morran, L. T. & Phillips, P. C. Males, Outcrossing, and Sexual Selection in Caenorhabditis Nematodes. *Genetics* **213**, 27–57 (2019).
9. Gimond, C. *et al.* Outbreeding depression with low genetic variation in selfing Caenorhabditis nematodes. *Evolution* **67**, 3087–3101 (2013).
10. C. elegans Sequencing Consortium. Genome sequence of the nematode C. elegans: a platform for investigating biology. *Science* **282**, 2012–2018 (1998).
11. Price, A. L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904–909 (2006).
12. Crombie, T. A. *et al.* Deep sampling of Hawaiian Caenorhabditis elegans reveals high genetic diversity and admixture with global populations. *Elife* **8**, (2019).
13. Andersen, E. C. *et al.* Chromosome-scale selective sweeps shape Caenorhabditis elegans genomic diversity. *Nat. Genet.* **44**, 285–290 (2012).

14. Rockman, M. V. & Kruglyak, L. Recombinational landscape and population genomics of Caenorhabditis elegans. *PLoS Genet.* **5**, e1000419 (2009).
15. Rockman, M. V., Skrovanek, S. S. & Kruglyak, L. Selection at linked sites shapes heritable phenotypic variation in C. elegans. *Science* **330**, 372–376 (2010).
16. Cutter, A. D. & Payseur, B. A. Genomic signatures of selection at linked sites: unifying the disparity among species. *Nat. Rev. Genet.* **14**, 262–274 (2013).
17. Kim, C. *et al.* Long-read sequencing reveals intra-species tolerance of substantial structural variations and new subtelomere formation in C. elegans. *Genome Res.* **29**, 1023–1035 (2019).
18. Thompson, O. A. *et al.* Remarkably Divergent Regions Punctuate the Genome Assembly of the Caenorhabditis elegans Hawaiian Strain CB4856. *Genetics* **200**, 975–989 (2015).
19. Greene, J. S. *et al.* Balancing selection shapes density-dependent foraging behaviour. *Nature* **539**, 254–258 (2016).
20. Koenig, D. *et al.* Long-term balancing selection drives evolution of immunity genes in Capsella. *Elife* **8**, (2019).
21. Barrett, R. D. H. & Schluter, D. Adaptation from standing genetic variation. *Trends Ecol. Evol.* **23**, 38–44 (2008).
22. Schulenburg, H., Hoeppner, M. P., Weiner, J., 3rd & Bornberg-Bauer, E. Specificity of the innate immune system and diversity of C-type lectin domain (CTLD) proteins in the nematode Caenorhabditis elegans. *Immunobiology* **213**, 237–250 (2008).
23. Reddy, K. C. *et al.* An Intracellular Pathogen Response Pathway Promotes Proteostasis in C. elegans. *Curr. Biol.* **27**, 3544–3553.e5 (2017).
24. van Sluijs, L. *et al.* Balancing selection shapes the Intracellular Pathogen Response in natural Caenorhabditis elegans populations. *bioRxiv* 579151 (2019) doi:10.1101/579151.
25. Bakowski, M. A. *et al.* Ubiquitin-mediated response to microsporidia and virus infection in C. elegans. *PLoS Pathog.* **10**, e1004200 (2014).
26. Chang, H. C., Paek, J. & Kim, D. H. Natural polymorphisms in C. elegans HECW-1 E3 ligase affect pathogen avoidance behaviour. *Nature* **480**, 525–529 (2011).
27. Troemel, E. R., Félix, M.-A., Whiteman, N. K., Barrière, A. & Ausubel, F. M. Microsporidia are natural

intracellular parasites of the nematode Caenorhabditis elegans. *PLoS Biol.* **6**, 2736–2752 (2008).

28. Félix, M.-A. *et al.* Natural and experimental infection of Caenorhabditis nematodes by novel viruses related to nodaviruses. *PLoS Biol.* **9**, e1000586 (2011).
29. Chen, K., Franz, C. J., Jiang, H., Jiang, Y. & Wang, D. An evolutionarily conserved transcriptional response to viral infection in Caenorhabditis nematodes. *BMC Genomics* **18**, 303 (2017).
30. Balla, K. M., Andersen, E. C., Kruglyak, L. & Troemel, E. R. A wild C. elegans strain has enhanced epithelial immunity to a natural microsporidian parasite. *PLoS Pathog.* **11**, e1004583 (2015).
31. Ashe, A. *et al.* A deletion polymorphism in the *Caenorhabditis elegans* RIG-I homolog disables viral RNA dicing and antiviral immunity. *Elife* **2**, e00994 (2013).
32. Martin, N., Singh, J. & Aballay, A. Natural Genetic Variation in the Caenorhabditis elegans Response to Pseudomonas aeruginosa. *G3* **7**, 1137–1147 (2017).
33. Samuel, B. S., Rowedder, H., Braendle, C., Félix, M.-A. & Ruvkun, G. Caenorhabditis elegans responses to bacteria from its natural habitats. *Proc. Natl. Acad. Sci. U. S. A.* **113**, E3941–9 (2016).
34. Seidel, H. S., Rockman, M. V. & Kruglyak, L. Widespread Genetic Incompatibility in *C. Elegans* Maintained by Balancing Selection. *Science* **319**, 589–594 (2008).
35. Thomas, C. G. *et al.* Full-genome evolutionary histories of selfing, splitting, and selection in Caenorhabditis. *Genome Res.* **25**, 667–678 (2015).
36. Cutter, A. D., Wasmuth, J. D. & Washington, N. L. Patterns of molecular evolution in Caenorhabditis preclude ancient origins of selfing. *Genetics* **178**, 2093–2104 (2008).
37. Stevens, L. *et al.* The Genome of Caenorhabditis bovis. *Curr. Biol.* **30**, 1023–1031.e4 (2020).
38. Félix, M.-A. & Duveau, F. Population dynamics and habitat sharing of natural populations of Caenorhabditis elegans and C. briggsae. *BMC Biol.* **10**, 59 (2012).
39. Kiontke, K. C. *et al.* A phylogeny and molecular barcodes for Caenorhabditis, with numerous new species from rotting fruits. *BMC Evol. Biol.* **11**, 339 (2011).
40. Ferrari, C. *et al.* Ephemeral-habitat colonization and neotropical species richness of Caenorhabditis nematodes. *BMC Ecol.* **17**, 43 (2017).
41. Schulenburg, H. & Félix, M.-A. The Natural Biotic Environment of Caenorhabditis elegans. *Genetics* **206**, 55–86 (2017).

42. Greene, J. S., Dobosiewicz, M., Butcher, R. A., McGrath, P. T. & Bargmann, C. I. Regulatory changes in two chemoreceptor genes contribute to a Caenorhabditis elegans QTL for foraging behavior. *Elife* **5**, (2016).

43. Ghosh, R., Andersen, E. C., Shapiro, J. A., Gerke, J. P. & Kruglyak, L. Natural variation in a chloride channel subunit confers avermectin resistance in C. elegans. *Science* **335**, 574–578 (2012).

44. Ben-David, E., Burga, A. & Kruglyak, L. A maternal-effect selfish genetic element in Caenorhabditis elegans. *Science* **356**, 1051–1055 (2017).

45. Cutter, A. D., Baird, S. E. & Charlesworth, D. High nucleotide polymorphism and rapid decay of linkage disequilibrium in wild populations of Caenorhabditis remanei. *Genetics* **174**, 901–913 (2006).

46. Dey, A., Chan, C. K. W., Thomas, C. G. & Cutter, A. D. Molecular hyperdiversity defines populations of the nematode Caenorhabditis brenneri. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 11056–11060 (2013).

47. Brandvain, Y., Slotte, T., Hazzouri, K. M., Wright, S. I. & Coop, G. Genomic identification of founding haplotypes reveals the history of the selfing species Capsella rubella. *PLoS Genet.* **9**, e1003754 (2013).

48. Nordborg, M., Charlesworth, B. & Charlesworth, D. Increased levels of polymorphism surrounding selectively maintained sites in highly selling species. *Proceedings of the Royal Society of London. Series B: Biological Sciences* **263**, 1033–1039 (1996).

49. Wiuf, C., Zhao, K., Innan, H. & Nordborg, M. The probability and chromosomal extent of trans-specific polymorphism. *Genetics* **168**, 2363–2372 (2004).

50. Todesco, M. *et al.* Massive haplotypes underlie ecotypic differentiation in sunflowers. *Nature* (2020) doi:10.1038/s41586-020-2467-6.

51. Burgarella, C. *et al.* Adaptive Introgression: An Untapped Evolutionary Mechanism for Crop Adaptation. *Front. Plant Sci.* **10**, 4 (2019).

52. Stebbins, G. L. Self Fertilization and Population Variability in the Higher Plants. *Am. Nat.* **91**, 337–354 (1957).

53. Andersen, E. C., Bloom, J. S., Gerke, J. P. & Kruglyak, L. A variant in the neuropeptide receptor npr-1 is a major determinant of Caenorhabditis elegans growth and physiology. *PLoS Genet.* **10**, e1004156 (2014).

54. Cook, D. E., Zdraljevic, S., Roberts, J. P. & Andersen, E. C. CeNDR, the Caenorhabditis elegans natural diversity resource. *Nucleic Acids Res.* **45**, D650–D657 (2017).

55. Cook, D. E. *et al.* The Genetic Basis of Natural Variation in Caenorhabditis elegans Telomere Length. *Genetics* **204**, 371–383 (2016).

56. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).

57. Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27**, 2987–2993 (2011).

58. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).

59. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv [q-bio.GN]* (2013).

60. Lee, R. Y. N. *et al.* WormBase 2017: molting into a new stage. *Nucleic Acids Res.* **46**, D869–D874 (2018).

61. Poplin, R. *et al.* Scaling accurate genetic variant discovery to tens of thousands of samples. *bioRxiv* 201178 (2018) doi:10.1101/201178.

62. Cingolani, P. *et al.* A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly* **6**, 80–92 (2012).

63. Ortiz, E. M. *vcf2phylip*. (Github).

64. Schliep, K. P. phangorn: phylogenetic analysis in R. *Bioinformatics* **27**, 592–593 (2011).

65. Yu, G., Smith, D. K., Zhu, H., Guan, Y. & Lam, T. T.-Y. ggtree : an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods Ecol. Evol.* **8**, 28–36 (2017).

66. Patterson, N., Price, A. L. & Reich, D. Population structure and eigenanalysis. *PLoS Genet.* **2**, e190 (2006).

67. Browning, B. L. & Browning, S. R. Detecting identity by descent and estimating genotype error rates in sequence data. *Am. J. Hum. Genet.* **93**, 840–851 (2013).

68. Miles, A., Ralph, P., Rae, S. & Pisupati, R. *cggh/scikit-allel: v1.2.1*. (2019). doi:10.5281/zenodo.3238280.

69. Zhang, C., Dong, S.-S., Xu, J.-Y., He, W.-M. & Yang, T.-L. PopLDdecay: a fast and effective tool for linkage disequilibrium decay analysis based on variant call format files. *Bioinformatics* **35**, 1786–1788 (2019).

70. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).

71. Pedersen, B. S. & Quinlan, A. R. Mosdepth: quick coverage calculation for genomes and exomes. *Bioinformatics* **34**, 867–868 (2018).

72. Holdorf, A. D. *et al.* WormCat: An Online Tool for Annotation and Visualization of Caenorhabditis elegans Genome-Scale Data. *Genetics* (2019) doi:10.1534/genetics.119.302919.

73. Yu, G., Wang, L.-G., Han, Y. & He, Q.-Y. clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS* **16**, 284–287 (2012).

74. Carlson, M. org.Ce.eg.db: Genome wide annotation for Worm. R package version 3.8.2. *Bioconductor* https://bioconductor.org/packages/release/data/annotation/html/org.Ce.eg.db.html (2019).

75. García-González, A. P. *et al.* Bacterial Metabolism Affects the C. elegans Response to Cancer Chemotherapeutics. *Cell* **169**, 431–441.e8 (2017).

76. Andersen, E. C. *et al.* A Powerful New Quantitative Genetics Platform, Combining *Caenorhabditis elegans* High-Throughput Fitness Assays with a Large Collection of Recombinant Strains. *G3* **5**, g3.115.017178–920 (2015).

77. Boyd, W. A., Smith, M. V. & Freedman, J. H. Caenorhabditis elegans as a model in developmental toxicology. *Methods Mol. Biol.* **889**, 15–24 (2012).

78. Shimko, T. C. & Andersen, E. C. COPASutils: an R package for reading, processing, and visualizing data from COPAS large-particle flow cytometers. *PLoS One* **9**, e111090 (2014).

79. *easysorter*. (Github).

80. Endelman, J. B. Ridge Regression and Other Kernels for Genomic Selection with R Package rrBLUP. *The Plant Genome Journal* **4**, 250–256 (2011).

81. Qiu, Y. *RSpectra*. (Github).

82. Bilgrau, A. E. *correlateR*. (Github, 2018).

83. Li, J. & Ji, L. Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix. *Heredity* **95**, 221–227 (2005).

84. Zdraljevic, S. *et al.* Natural variation in C. elegans arsenic toxicity is explained by differences in branched chain amino acid metabolism. *Elife* **8**, (2019).

85. Riemondy, K. A. *et al.* valr: Reproducible genome interval analysis in R. *F1000Res.* **6**, 1025 (2017).

86. Team, R. C. R: A language and environment for statistical computing [Internet]. Vienna, Austria; 2014. (2017).

87. Ruan, J. & Li, H. Fast and accurate long-read assembly with wtdbg2. *Nat. Methods* **17**, 155–158 (2020).

88. Kolmogorov, M., Yuan, J., Lin, Y. & Pevzner, P. A. Assembly of long, error-prone reads using repeat graphs. *Nat. Biotechnol.* **37**, 540–546 (2019).

89. Koren, S. *et al.* Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res.* **27**, 722–736 (2017).

90. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015).

91. Walker, B. J. *et al.* Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* **9**, e112963 (2014).

92. Laetsch, D. R. & Blaxter, M. L. BlobTools: Interrogation of genome assemblies. *F1000Res.* **6**, 1287 (2017).

93. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).

94. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

95. Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).

96. Pundir, S., Martin, M. J. & O'Donovan, C. UniProt Protein Knowledgebase. in *Protein Bioinformatics: From Protein Modifications and Networks to Proteomics* (eds. Wu, C. H., Arighi, C. N. & Ross, K. E.) 41–55 (Springer New York, 2017).

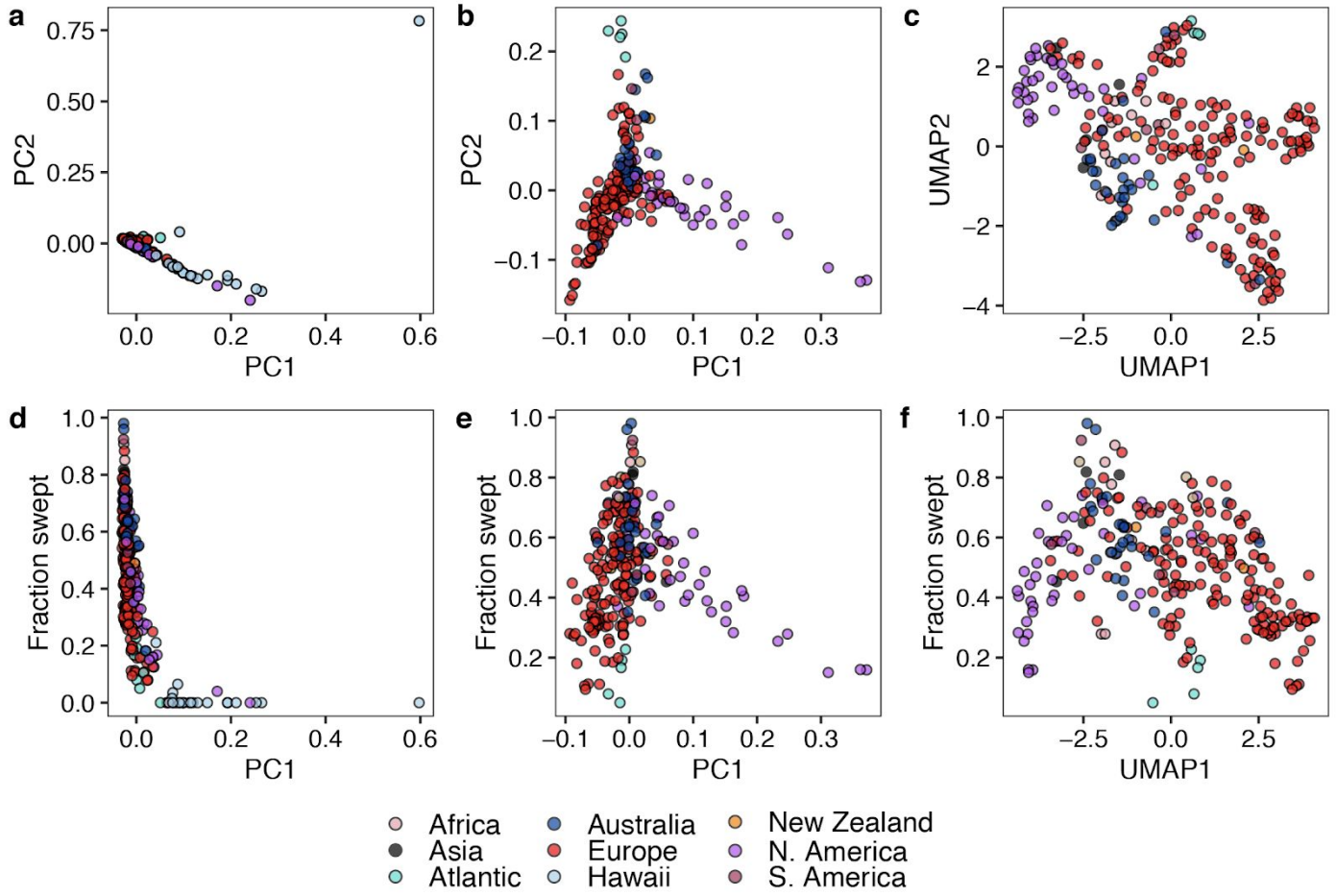
97. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat.*

*Methods* **12**, 59–60 (2015).

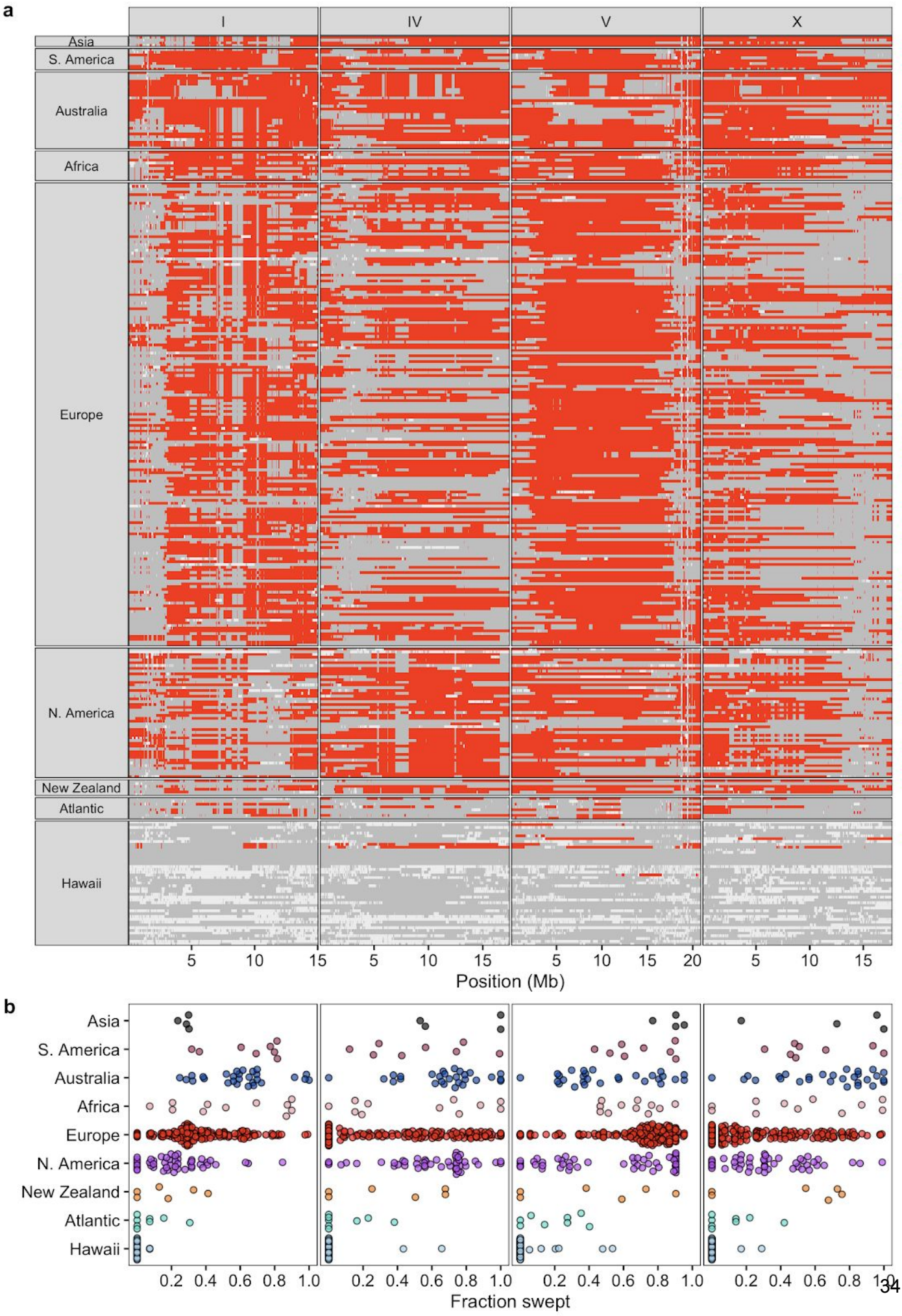
98. R., S. A. H. RepeatMasker. http://www.repeatmasker.org (2008-2015).
99. Bao, W., Kojima, K. K. & Kohany, O. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob. DNA* **6**, 11 (2015).
100. Stanke, M., Diekhans, M., Baertsch, R. & Haussler, D. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics* **24**, 637–644 (2008).
101. Consortium, T. C. E. S. Genome Sequence of the Nematode C. elegans: A Platform for Investigating Biology. *Science* 1–8 (1998).
102. Delcher, A. L., Salzberg, S. L. & Phillippy, A. M. Using MUMmer to identify similar regions in large sequence sets. *Curr. Protoc. Bioinformatics* **Chapter 10**, Unit 10.3 (2003).
103. Emms, D. M. & Kelly, S. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol.* **16**, 157 (2015).
104. Slater, G. S. C. & Birney, E. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* **6**, 31 (2005).
105. Finn, R. D. *et al.* The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res.* **44**, D279–85 (2016).
106. Jones, P. *et al.* InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240 (2014).
107. Wickham, H. ggplot2: elegant graphics for data analysis. (2016).
108. Bradley, R. K. *et al.* Fast statistical alignment. *PLoS Comput. Biol.* **5**, e1000392 (2009).
109. Minh, B. Q. *et al.* IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Mol. Biol. Evol.* **37**, 1530–1534 (2020).
110. Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A. & Jermiin, L. S. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods* **14**, 587–589 (2017).
111. Zheng, X. *et al.* A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics* **28**, 3326–3328 (2012).
112. Stein, L. D. *et al.* The genome sequence of Caenorhabditis briggsae: a platform for comparative genomics. *PLoS Biol.* **1**, E45 (2003).

113. Yin, D. *et al.* Rapid genome shrinkage in a self-fertile nematode reveals sperm competition proteins. *Science* **359**, 55–61 (2018).

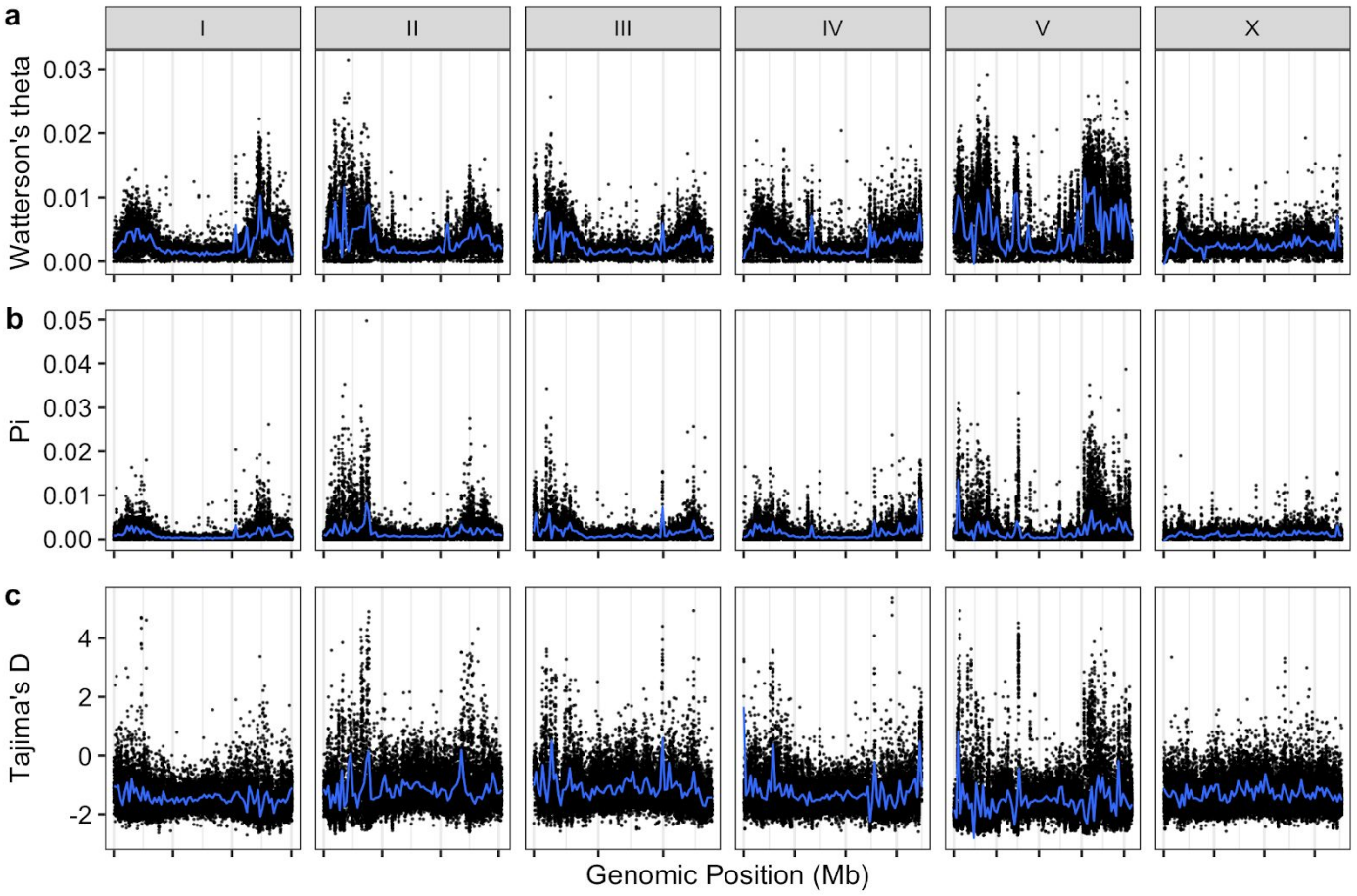
## Extended data figures



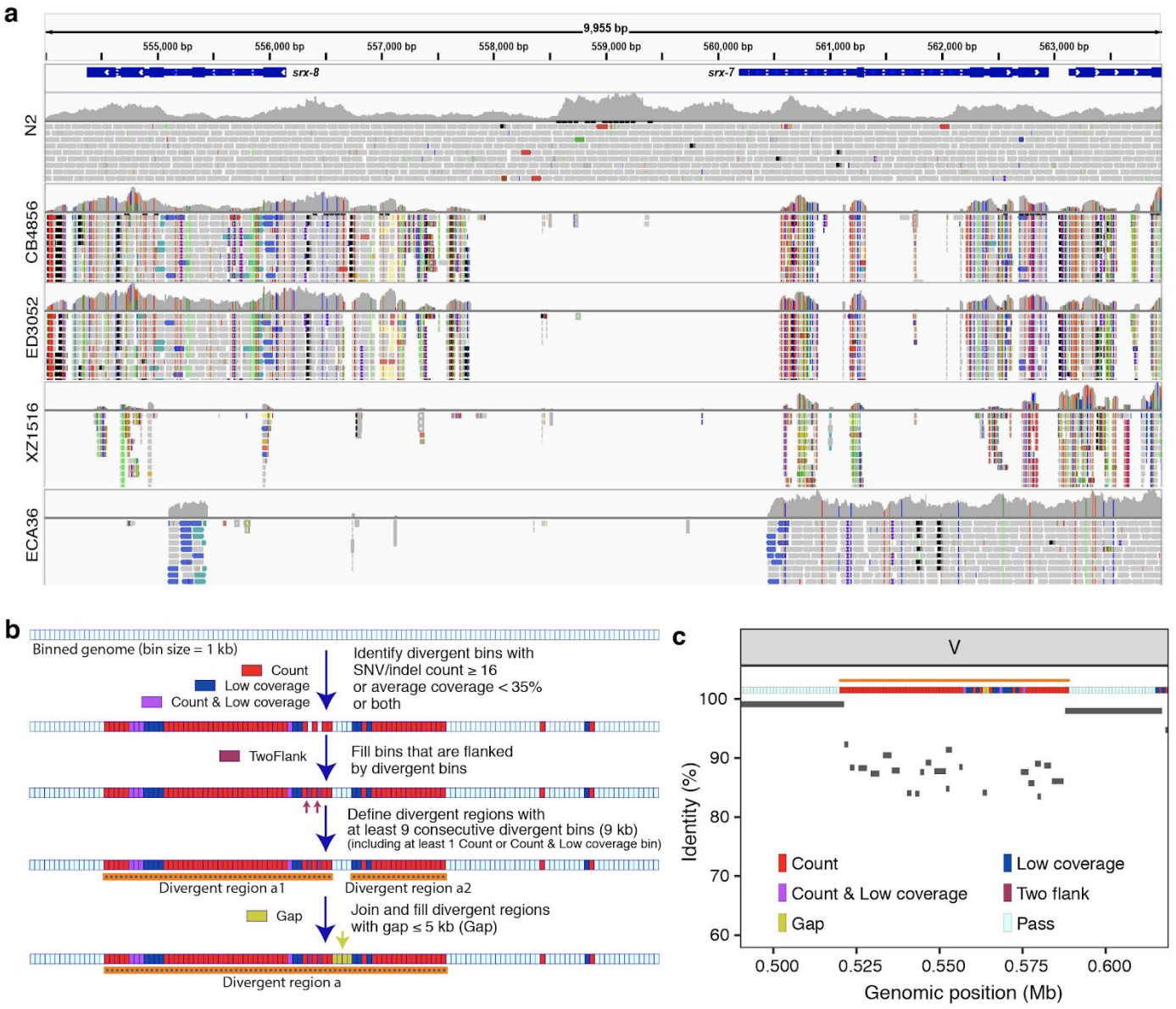

**Extended Data Fig. 1 | Selective sweeps shape population structure of the Global group**
(a) Plot of 308 isotypes that belong to the Global group according to their values for each of the two significant axes of variation, as determined by PCA of the genotype covariances. Each point is one of the 308 isotypes, which is colored by their geographic origins.
(b) Plot of 251 isotypes from the Global group according to the values for each of the two significant axes of variation, as determined by PCA of the genotype covariances with five iterations of outlier removal. Each point is one of the 251 isotypes, which is colored by the geographic origin.
(c) Plot of 251 isotypes from the Global group according to the values for each of the two significant axes of variation, as determined by Uniform Manifold Approximation and Projection (UMAP) analysis using five principal components from PCA of the genotype covariances with five iterations of outlier removal. Each point is one of the 251 isotypes, which is colored by the geographic origin.
(d-f) Scatter plots for the fraction of swept chromosomes for the isotypes in (a-c), respectively. The first significant axis of variation defined in (a-c) is shown on the x-axis, and the fraction of swept chromosomes is shown on the y-axis. Each point corresponds to one of 308 (d) or 251 (e,f) isotypes and is colored by the geographic origin.



**Extended Data Fig. 2 | Chromosome-scale selective sweeps across wild *C. elegans* isotypes**
(a) Sharing of the swept haplotype (red) among 324 wild isotypes with known geographic origin is shown. Gray genomic regions represent non-swept haplotypes, and white represents unclassified haplotypes. Each row is one of the 324 isotypes, grouped by the geographic origin. The genomic position in Mb is plotted on the x-axis, and each tick mark represents 5 Mb of the chromosome. Only swept chromosomes (I, IV, V, and X) are shown (Methods).
(b) Beeswarm plots of the proportion of swept haplotype for each chromosome from (a) for 324 isotypes with known geographic origins are shown. Wild isotypes are grouped by geographic origin. Each point corresponds to one of the 324 isotypes, and geographic origins are shown on the y-axis.




**Extended Data Fig. 3 | Patterns of molecular diversity across the *C. elegans* genome**
Chromosomal patterns of genetic diversity statistics are shown. Sliding window analyses of (a) Watterson's theta, (b) nucleotide diversity (pi), and (c) Tajima's D were performed with a sliding window of size 1 kb and a step size of 1 kb. Each dot corresponds to the calculated value for a particular window. The genomic position in Mb is plotted on the x-axis, and each tick represents 5 Mb of the chromosome. Diversity statistic values are shown on the y-axis. Smoothed lines (blue) are LOESS fits.

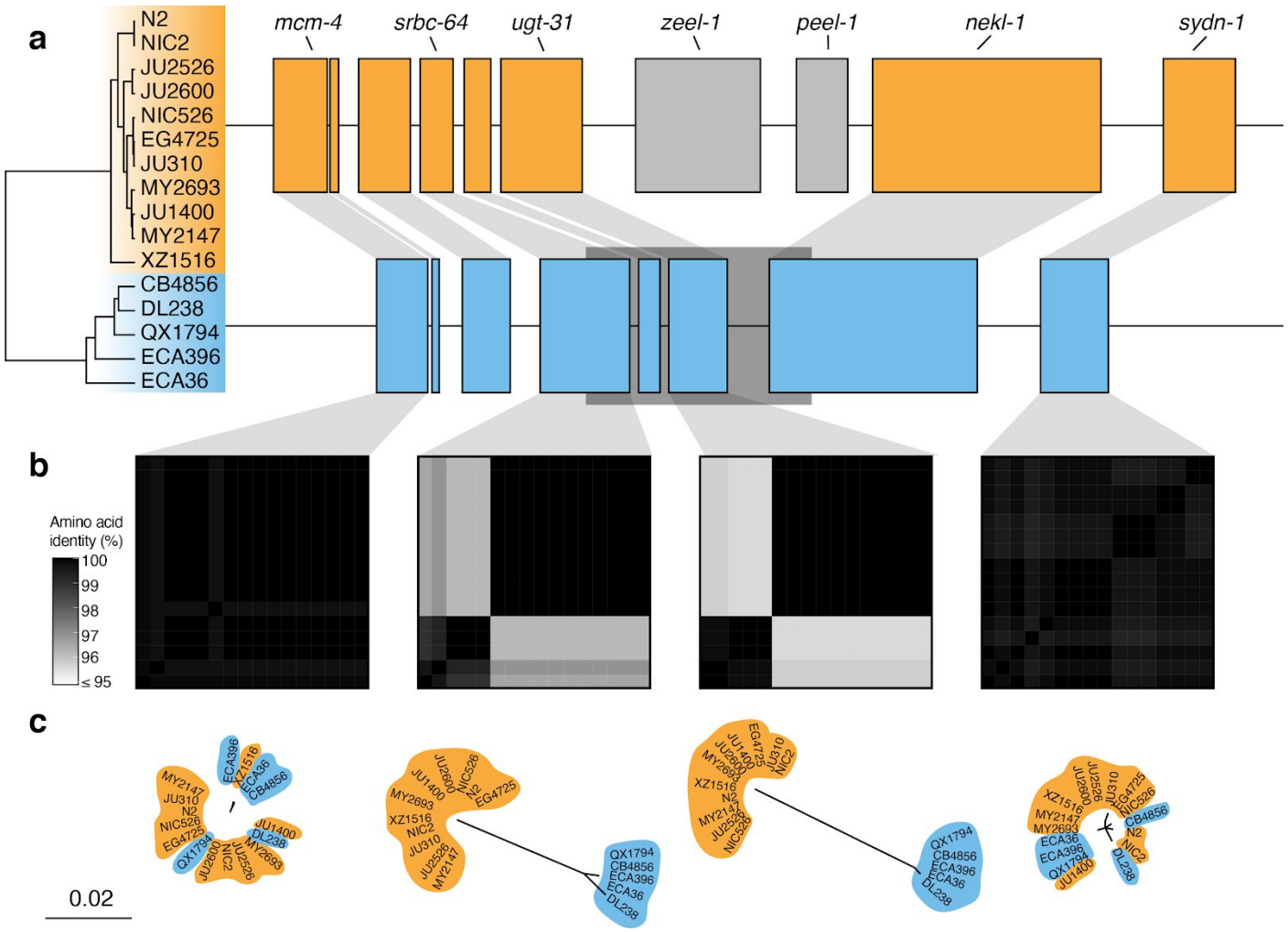



**Extended Data Fig. 4 | Characterization of hyper-divergent regions at the isotype level**

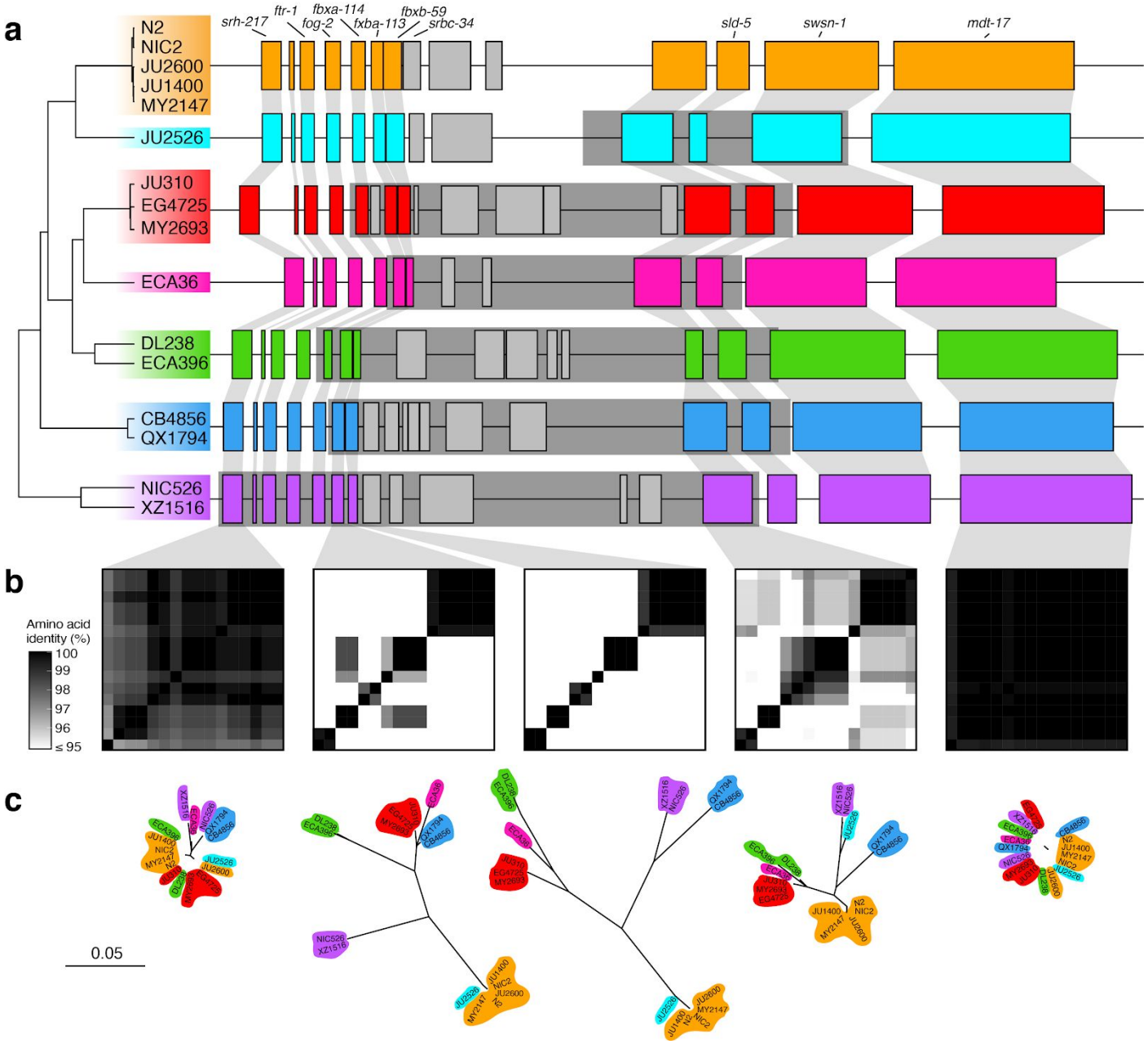
(a) Short-read alignments from five isotypes (N2, CB4856, ED3052, XZ1516, and ECA36) to the N2 reference genome (WS245) for a region of chromosome V (V:554,000-564,000) plotted by IGV 2.8.x are shown. Genes (*srx-7* and *srx-8*) in the interval are shown at the top. For each isotype, the top panel shows the coverage of genomic positions and the bottom panel shows aligned short-reads at genomic positions (gray: normal reads, red: reads with putative deletion, blue: reads with putative insertion, navy and turquoise: reads with putative inversion, green: reads with putative duplication or translocation). Colored vertical lines indicate mismatched bases at the position.

(b) The workflow for the characterization of hyper-divergent regions at the isotype level is shown (See Methods).

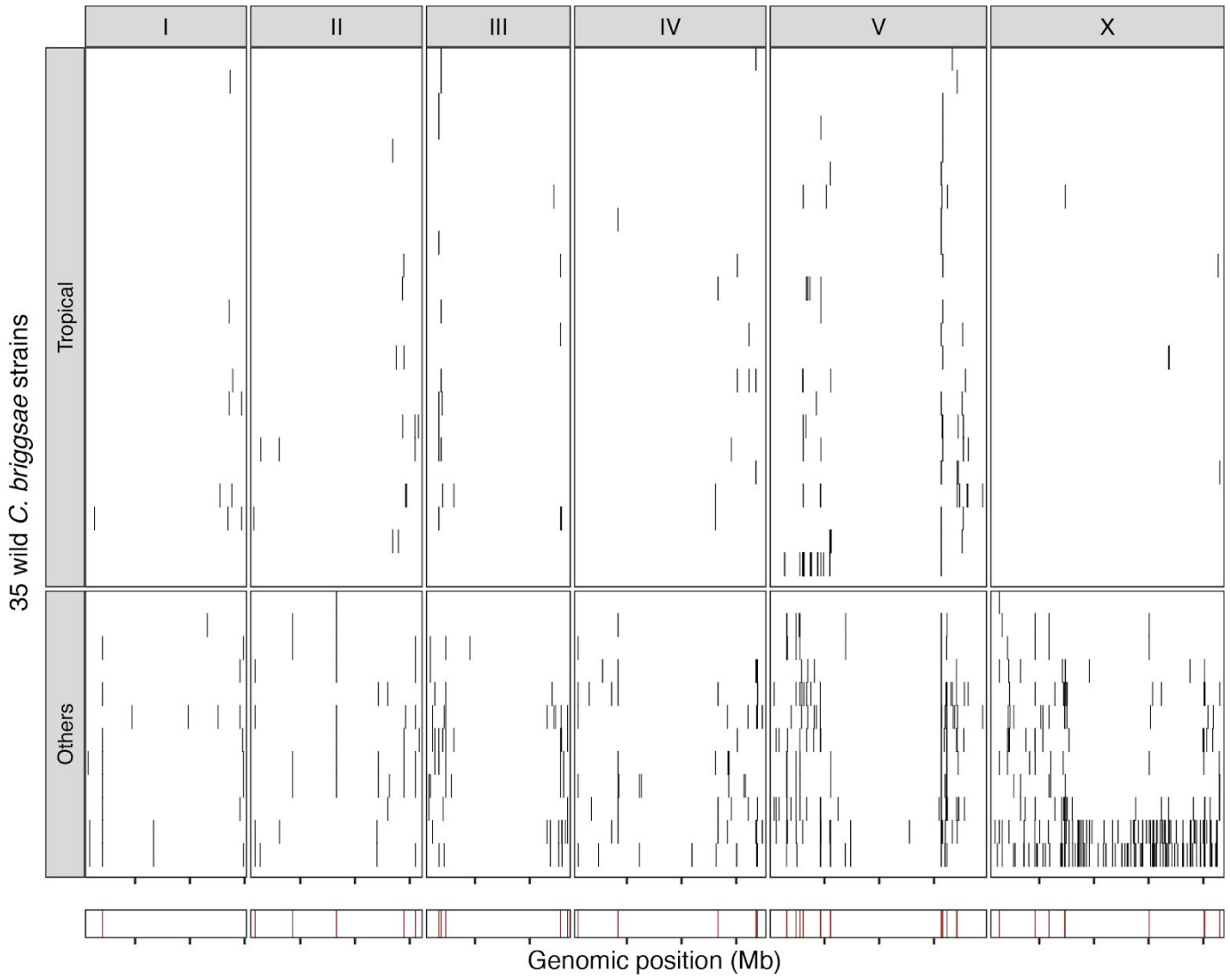
(c) A plot for the validation of short-read based characterization of a hyper-divergent region with the long-read alignment is shown. The orange horizontal line denotes a hyper-divergent region (V:520,000-589,000) in the locus (V:490,000-619,000) of CB4856. The multi-colored horizontal rectangle shows classifications of genomic bins in the locus. Gray bars correspond to the alignments from CB4856 long-read sequences to the N2 reference genome (WS245), and identities of alignments are shown on the y-axis.




**Extended Data Fig. 5 | Two hyper-divergent haplotypes at the *peel-1 zeel-1* incompatibility locus**
(a) The protein-coding gene contents of the two hyper-divergent haplotypes at the *peel-1 zeel-1* incompatibility locus on the left arm of chromosome I (I:2,318,291-2,381,851 of the N2 reference genome). The tree was inferred using SNVs and coloured by inferred haplotypes. For each distinct haplotype, we chose a single isotype as a haplotype representative (orange haplotype: N2, blue haplotype: CB4856) and predicted protein-coding genes using both protein-based alignments and *ab initio* approaches. Protein-coding genes are shown as boxes; those genes that are conserved in all haplotypes are coloured based on their haplotype, and those genes that are not are coloured light gray. Dark grey boxes behind genes indicate coordinates of divergent regions. Genes with locus names in N2 are highlighted.
(b) Heatmaps showing amino acid identity for alleles of four genes (*mcm-4*, *srbc-64*, *ugt-31*, and *sydn-1*). The percentage identity was calculated using alignments of protein sequences from all 16 isotypes. Heatmaps are ordered by the SNV tree shown in (a).
(c) Maximum-likelihood gene trees of four genes (*mcm-4*, *srbc-64*, *ugt-31,* and *sydn-1*) inferred using amino acid alignments. Trees are plotted on the same scale (scale shown; scale is in substitutions per site). Strain names are coloured by their haplotype.




**Extended Data Fig. 6 | Hyper-divergent haplotypes at a region on the right arm of chromosome V**
(a) The protein-coding gene contents of the seven hyper-divergent haplotypes at a region on the right arm of chromosome V (V:20,193,463-20,267,244 of the N2 reference genome). The tree was inferred using SNVs and coloured by inferred haplotypes. For each distinct haplotype, we chose a single isotype as a haplotype representative (orange haplotype: N2, light blue haplotype: JU2526, red haplotype: EG4725, pink haplotype: ECA36, green haplotype: DL238, dark blue haplotype: QX1794, purple haplotype: NIC526) and predicted protein-coding genes using both protein-based alignments and *ab initio* approaches. JU2526 shares the reference haplotype at *fbxa-113* and *fbxb-59* (six hyper-divergent haplotypes at these loci) but is divergent at *Y113G7B.15* (seven hyper-divergent haplotypes at this locus). Protein-coding genes are shown as boxes; those genes that are conserved in all haplotypes are coloured based on their haplotypes, and those genes that are not are coloured light gray. Dark grey boxes behind genes indicate coordinates of divergent regions. Genes with locus names in N2 are highlighted.
(b) Heatmaps showing amino acid identity for between alleles of five genes (*srh-217*, *fbxb-113*, *fbxb-59*, *Y113G7B.15*, and *mdt-17*). The percentage identity was calculated using alignments of proteins sequences from all 16 isotypes. Heatmaps are ordered by the SNV tree shown in (a).
(c) Maximum-likelihood gene trees of five genes (*srh-217*, *fbxb-113*, *fbxb-59*, *Y113G7B.15*, and *mdt-17*) inferred using amino acid alignments. Trees are plotted on the same scale (scale shown; scale is in substitutions per site). Strain names are coloured by their haplotype.




**Extended Data Fig. 7 | Hyper-divergent regions in *C. briggsae***

The genome-wide distribution of hyper-divergent regions across 35 non-reference wild *C. briggsae* strains is shown. In the top panel, each row is one of the 35 strains, grouped by previously defined clades (tropical or others) ordered by the total amount of genome covered by hyper-divergent regions (black). In the bottom panel, brown bars indicate genomic positions in which more than 10% of strains are classified as hyper-divergent at the locus. The genomic position in Mb is plotted on the x-axis, and each tick represents 5 Mb of the chromosome.