Molecular Biology and Evolution, 2026, 43, 1–23
https://doi.org/10.1093/molbev/msag202
Published: 12 August 2026
Discoveries

Population structure and punctuated genomic hyper-diversity in *Caenorhabditis briggsae*

Nicolas D. Moya [1], Bowen Wang [1], Robyn E. Tanny [1], Michael E.G. Sauria [1], Lance M. O'Connor [1], Ayeh Khorshidian[1], Ryan McKeown [2], Charlie Gosse [3], Clayton M. Dilks [2], Timothy A. Crombie [2,4], Gaotian Zhang [2,3], Emha Rais[5], Lise Frézal [3], Viet Dai Dang [6,7], Elkana Haryoso[5], Mia P. Devi [5], Clotilde Gimond [8], Daniel E. Cook [2], Jung-Chen Hsu [6], Amanda O. Shaver [1], Stefan Zdraljevic [2], Aurélien Richaud[3], Tongshu Wen[9], Aatira Mehraj[10], Sharanya H [10], Karthick Raja Arulprakasam [10], Emily J. Koury [2], Nicole M. Roberto[2], Etta S. Schaye [1], Varsha Singh [10,11], Hagus Tarno [5], Michael Ailion [12], Annalise B. Paaby [13], Zhongying Zhao [9], Asher D. Cutter [14], John Wang [6], Matthew V. Rockman [15], Marie-Anne Félix [3], Christian Braendle [8], Erik C. Andersen [1]*

[1]Department of Biology, Johns Hopkins University, Baltimore, MD, USA
[2]Department of Molecular Biosciences, Northwestern University, Evanston, IL, USA
[3]Institut de Biologie de l'Ecole Normale Supérieure, CNRS, Inserm, Paris, France
[4]Department of Biomedical and Chemical Engineering and Sciences, Florida Institute of Technology, Melbourne, FL, USA
[5]Department of Pests and Disease, Faculty of Agriculture, University of Brawijaya, Malang, East Java, Indonesia
[6]Biodiversity Research Center, Academia Sinica, Taipei, Taiwan
[7]Southern Institute of Ecology, Institute of Applied Material Science, Vietnam Academy of Science and Technology, Ho Chi Minh, Vietnam
[8]Université Côte d'Azur, CNRS, Inserm, IBV, Nice, France
[9]Department of Biology, Hong Kong Baptist University, Hong Kong, China
[10]Indian Institute of Science, Bengaluru, India
[11]Division of Molecular Microbiology, School of Life Sciences, University of Dundee, Dundee, Scotland, UK
[12]Department of Biochemistry, University of Washington, Seattle, WA, USA
[13]School of Biological Sciences, Georgia Institute of Technology, Atlanta, GA, USA
[14]Department of Ecology & Evolutionary Biology, University of Toronto, Toronto, Ontario, Canada
[15]Department of Biology and Center for Genomics & Systems Biology, New York University, New York, NY, USA
***Corresponding author:** E-mail: erik.andersen@gmail.com.
Associate editor: Ilya Ruvinsky

Abstract

Comparative genomics provides a powerful framework to uncover the molecular and evolutionary mechanisms that shape genetic diversity, revealing how shared or lineage-specific processes influence their evolutionary trajectories. The nematode *Caenorhabditis briggsae* is distributed world-wide and is a comparative model to *Caenorhabditis elegans* in the biology of development, cellular mechanisms, neurobiology, complex trait mappings, and evolution. Following massive collection efforts by the nematode research community, we present the isolation of over 1,900 wild strains and analyses of genome sequences that catalog over six million single-nucleotide and insertion–deletion variants. These resources provide a powerful means to interrogate the causal genetic bases of phenotypic variation. Additionally, we describe *C. briggsae* population structure and discover new, genetically distinct groups within this primarily self-fertilizing species, including groups of highly related strains sampled across entire continents. We leveraged expansive genetic variation to decipher the effects of linkage and selection on the distribution of genetic diversity across the genome and across geographic regions. Within the species, we find genomic regions with extremely high levels of genetic variation similar to hyper-divergent regions found in *C. elegans* and other species. These regions harbor new genes and variation enriched for environmental sensing and pathogen responses. Based on comparisons to the outbreeding sister species *Caenorhabditis nigoni*, we conclude that long-term balancing selection has maintained substantial functional variation, likely associated with ecological variation, within *C. briggsae* since its divergence from an outbreeding ancestor. Overall, this massive strain resource enables future comparative genetics studies, including genome-wide association study contrasts between *Caenorhabditis* species.

Keywords comparative genomics, population genetics, genetic diversity, evolutionary biology

Received: March 2, 2026. **Revised:** July 13, 2026. **Accepted:** July 23, 2026

Introduction

Comparative population genomic analyses among closely related species are invaluable for investigating the evolutionary forces that shape genome diversity both within and among species. Such cross-species population genomic studies reveal whether genetic mechanisms act universally across taxa or reflect lineage-specific phenomena, providing insights into how demographic and selective processes can influence both neutral and functional genetic variation (Cao et al. 2011; Machado et al. 2016; Koenig et al. 2019; Abrams et al. 2021; Yang et al. 2024). In some cases, closely related species that occupy similar niches exhibit convergent evolution in genotypes and phenotypes (Machado et al. 2016; Yang et al. 2024). For example, the model organism *Drosophila melanogaster* and its sister species *Drosophila simulans* have independently evolved parallel clinal adaptation in genes along climatic gradients to diverse environments (Machado et al. 2016). In other cases, shared genotypes among distinct species might be caused by the maintenance of ancestral alleles (Cao et al. 2011; Koenig et al. 2019). For example, long-term balancing selection has maintained ancient haplotypes across plant species in the *Capsella* genus, where ancestral polymorphism at immunity-related loci might have been maintained despite divergent pathogen pressures (Koenig et al. 2019). In modern humans, global variation in diverse traits can be traced to adaptive introgression of alleles from closely related but now-extinct archaic Neanderthal and Denisovan populations, including skin and hair pigmentation, metabolism, and immunity (Yan et al. 2021; Reilly et al. 2022; Li et al. 2024). Comparative analyses also can decipher the evolutionary origins of species-specific traits. For example, comparisons between the yeasts *Saccharomyces cerevisiae* and *Saccharomyces paradoxus* revealed that thermal-tolerance loci evolved under positive selection in an ancestral population of *S. cerevisiae* before spreading across diverse regions and ecological niches (Abrams et al. 2021). It is unresolved to what extent evolutionary convergence, maintenance of ancestral polymorphism, adaptive introgression, and adaptive divergence each contribute to the genetic basis of biodiversity in different taxa. We must use a comparative population genomics approach to answer this question.

Caenorhabditis nematodes offer a strong comparative platform for these types of studies. The nematode *Caenorhabditis briggsae* is a self-fertilizing species related to the widely adopted model species *Caenorhabditis elegans* (Baird and Chamberlin 2006; Coghlan et al. 2006; Hillier et al. 2007). Both species share similar ecology, naturally proliferating in rotting plant material and often isolated in the same substrate (Félix and Duveau 2012; Crombie et al. 2019). Additionally, both species have independently evolved self-fertilization from outbreeding ancestors (Cutter et al. 2008; Kiontke et al. 2011). Extensive collections of wild *C. elegans* strains have been used to establish the genetic basis of phenotypic variation across various traits and to study the effects of self-fertilization on the distribution of genetic diversity across their genomes (Andersen et al. 2012; Cook et al. 2016). Notably, a recent study revealed that *C. elegans* carries extreme genetic variation concentrated in punctuated genomic regions thought to be maintained by long-term balancing selection since the evolution of self-fertilization (Lee et al. 2021), despite experiencing a general reduction in genetic diversity relative to its outbreeding relatives (Cutter et al. 2006a; Teterina et al. 2023), largely because of chromosome-scale positive selective sweeps (Andersen et al. 2012) and background selection (Rockman et al. 2010). These hyper-divergent regions are enriched for genes associated with sensory perception and response to pathogens, suggesting that they contribute to adaptation and survival in diverse ecological circumstances.

The evolution of regions with extreme genetic variation associated with environmental responses is a recurrent pattern across diverse taxa in the tree of life. Protozoan parasites such as *Plasmodium falciparum* and *Trypanosoma brucei* maintain large repertoires of diverse genes encoding surface antigens that facilitate the evasion of host immune systems (Kyes et al. 2007; Stockdale et al. 2008). Extreme genetic variation has also been documented across plant self-incompatibility loci, such as in outcrossing crucifer species of *Arabidopsis* and *Brassica* genera (Kusaba et al. 2001) and across immunity genes in selfing *Capsella* species (Koenig et al. 2019). Moreover, highly diverse regions surrounding host interaction genes have also been recently identified in parasitic nematodes *Heligmosomoides bakeri*, *Strongyloides ratti*, and *Globodera rostochiensis* (Eves-van den Akker et al. 2016; Cole et al. 2023; Stevens et al. 2023). Among vertebrate species, the major histocompatibility complex maintains alleles with exceptional sequence diversity that enable recognition and response to diverse and rapidly evolving pathogens (Klein 1980; Takahata and Nei 1990). Across these systems, several mechanisms have been proposed to explain the origin and maintenance of such extreme diversity and that could, separately or together, explain the emergence of hyper-divergent regions (HDRs) across *Caenorhabditis* nematodes (Moya et al. 2025). Additionally, the generally low levels of genetic variation outside of HDRs provides a unique opportunity to establish connections between mechanisms of genetic variation across heterogeneous environments. Although *C. briggsae* offers a powerful comparative platform to study these conserved or convergent patterns of extreme genetic diversity with *C. elegans*, a genome-wide analysis has yet to be explored for a comparably large *C. briggsae* strain collection to identify and compare HDRs.

Early studies of multilocus genotypes across wild *C. briggsae* strains uncovered evidence of genetic differentiation among geographically separated groups, particularly between strains isolated from tropical and temperate latitudes (Graustein et al. 2002; Jovelin et al. 2003; Cutter et al. 2006b). The incorporation of strains from new geographic locations continued to support population structure across latitudinal ranges but also identified rare, geographically restricted divergent haplotypes (Dolgin et al. 2008; Félix et al. 2013). More recently, investigation of the *C. briggsae* evolutionary history using a population genomics study of 37 wild whole-genome sequences (Thomas et al. 2015) implicated a near-simultaneous split of at least six distinct “lineages” (which we will call relatedness groups in this study) approximately 200,000 generations ago preceded by a more ancient split from all other relatedness groups approximately two million generations ago. Importantly, recent efforts have generated a vastly improved *C. briggsae* reference genome using Oxford Nanopore sequencing and provided updated gene models from high-quality transcriptome data and extensive manual curation (Stevens et al. 2022; Moya et al. 2023). These advances have narrowed the genome resource gap between *C. elegans* and *C. briggsae*, priming a comprehensive analysis of *C. briggsae* collections to enable large-scale comparative analyses of *Caenorhabditis* species.

In this study, we sequenced the whole genomes of over 1,900 wild *C. briggsae* strains collected from all over the world by the *Caenorhabditis* research community, massively expanding previous *C. briggsae* population studies. To characterize natural variation across the species, we leveraged updated genomic resources to discover single-nucleotide and insertion–deletion variants to reveal extensive genetic and geographic differentiation, allowing us to classify genetically distinct strains into 12 relatedness groups that likely represent independently evolving populations. Notably, strains with near identical whole-genome haplotypes have been isolated widely across geographic regions, suggesting that *C. briggsae* has undergone recent human-assisted dispersal and expansion. The distribution of polymorphism across *C. briggsae* genomes mirrors that of *C. elegans* genomes, such that chromosomal arm domains with low gene density and high-recombination rates display greater diversity relative to gene-dense, low-recombination chromosomal center domains. Our investigation of the patterns of genetic variation also revealed widespread HDRs that contain 31% of single-nucleotide variants (SNVs) and cover 6% of individual genomes on average. Much like *C. elegans*, we identified numerous examples of HDRs that display remarkable gene content variation and are enriched for genes associated with environmental responses. We also generated high-quality genome and gene resources for 16 *C. nigoni* inbred lines to interrogate the evolutionary origins of HDRs. Using estimates of synonymous site divergence between *C. briggsae* relatedness groups and between *C. briggsae* and *C. nigoni*, we find that hyper-divergent haplotypes do not predate speciation, consistent with no detectable excess of shared alleles with *C. nigoni* within HDRs relative to the entire genome. We conclude that HDRs likely originated prior or near to the split of the most diverged *C. briggsae* relatedness groups and were retained possibly by balancing selection acting on isolated groups or by cross-breeding between groups. Overall, the updated genomic resources, catalog of genetic diversity, population structure analysis, and characterization of HDRs for this large collection of *C. briggsae* strains will improve power to detect associations between genotype and phenotype and to inform comparative and intra-specific models of evolution.

Results

Caenorhabditis briggsae genome-wide variation reveals both regional and localized phylogeographic structure

Caenorhabditis briggsae has a wide global distribution with samples collected across six continents and many oceanic islands (Fig. 1a and b; Table S1). These samples come from bacteria-laden substrates of rotting vegetal materials. We compared whole-genome sequence data from 1,972 wild *C. briggsae* strains to understand patterns and levels of genetic variation (see Methods). Because of self-fertilization, wild *C. briggsae* strains are naturally inbred and essentially homozygous at all loci, so we classified the 1,972 strains into 713 distinct genome-wide haplotypes (referred to as isotypes, see Methods) (Fig. S1, Table S1). The majority of isotypes are composed of a single strain ($n = 470$) or multiple strains isolated less than 1 km away from each other ($n = 173$). Two isotypes comprise multiple strains where all but one strain lack information of their geographic site of isolation. Additionally, we classified the remaining 68 isotypes as regional, composed of strains isolated in the same geographic region between 1 km and up to 5,000 km from each other (Fig. S1). We further investigated the genotype similarity and fine-scale geographic distribution of the largest regional isotype (NIC174). This isotype includes 261 strains sampled in Europe (Fig. S2). Within the NIC174 isotype, we compared the geographic distribution of strains against genotype identities at all segregating SNV sites and found that strains from the same locality often display increased genotype similarity (Fig. S3). This result suggests long-distance dispersal followed by recent local divergence. We applied a molecular clock at 4-fold degenerate sites and assumed a *C. elegans* mutation rate (Saxena et al. 2019) to estimate that approximately 1,000 generations separate the most genetically dissimilar strains within this isotype. Assuming a minimum of 10 generations per year, we estimated that this isotype diverged less than 100 years ago. This isotype is consistent with the rapid dispersal and establishment of a single genome-wide haplotype likely because of recent human activity, suggesting an unprecedented sweep even more extensive than chromosomal-level selective sweeps found among European strains in *C. elegans* (Andersen et al. 2012; Lee et al. 2021).

Across the genomes of the 713 isotype reference strains, we identified approximately 5.65 million SNVs and 878,376 short (<50 base pairs (bp)) insertion–deletion (indel) variants (see Data Availability, Methods). After pruning variants in strong linkage disequilibrium (LD) ($r^2 > 0.9$, see Methods), we used biallelic SNVs to perform principal component analysis (PCA) (Fig. 1c and d). Nearly all isotypes fall within a single cluster in principal component (PC) space, separated from the minority of isotypes that define discrete clusters and also come from narrow geographic regions (Fig. 1d). Four PCs distinguish a cluster of 32 isotypes from the east coast of Australia, a cluster of 10 isotypes from the south of India, a cluster of 26 isotypes from Taiwan, and six small clusters (1 to 7 isotypes) from various geographic regions (Indonesia, Taiwan, Hubei, Quebec) (Fig. 1d, Table S2). The two remaining clusters cover a wide geographic range; one cluster of four isotypes from Wisconsin, USA or Nairobi, Africa and a cluster of 27 isotypes isolated from temperate latitudes (Europe, North America, and Japan). Chromosome-specific PCs largely recapitulate the patterns observed genome-wide (Fig. S4), with the exception of divergent clusters from India and Australia occupying the same PC space when considering SNVs from chromosome II or X alone. We performed iterative outlier removal followed by PCA to further dissect genetic differentiation in the Global population cluster (Fig. S5) and found few distinct clusters, leading us to pursue additional approaches to explore *C. briggsae* population genetic differentiation.

We estimated mean pairwise genetic similarity by calculating the proportion of identical alleles across genome-wide SNVs between isotypes and clustered all isotypes into 12 relatedness groups (Fig. 2). These groups showed a mean within-group genetic similarity of >98% but <93% between-group genetic similarity across all identified SNV sites (Fig. S6). Eight of these groups represent geographically restricted PC clusters, including a group from Australia (referred to as Australia Divergent or AD group), a group from India (referred to as Kerala Divergent or

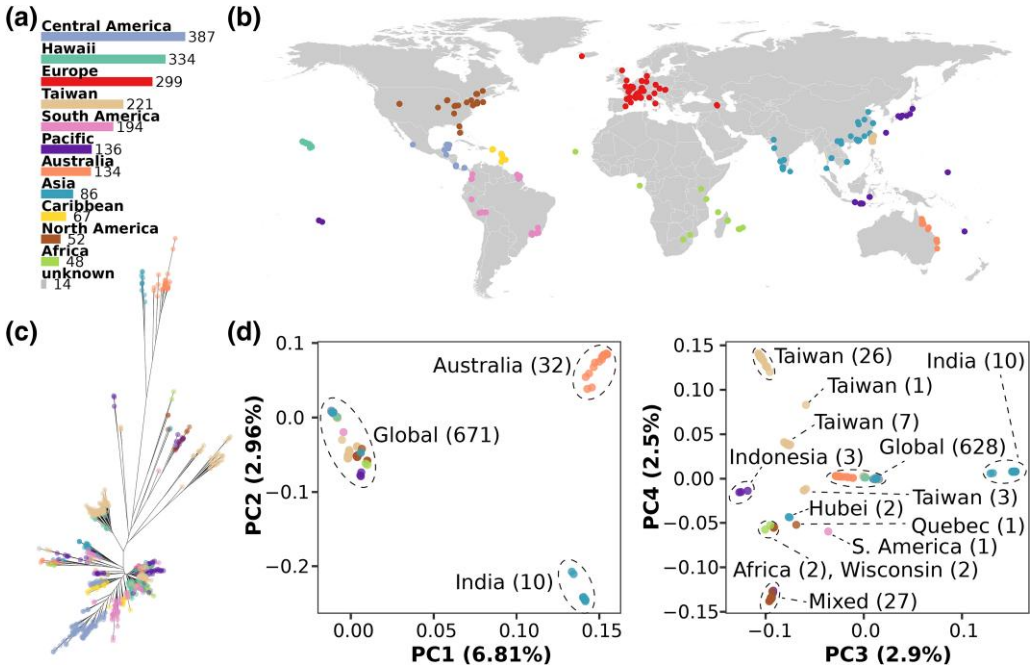

Figure 1 Global sampling efforts define 713 genetically distinct strains. a) Total number of strains in each geographic region. b) Global sampling map. c) Unrooted equal angle maximum likelihood tree of 713 *C. briggsae* isotype reference strains. The tree was generated from LD-pruned variants with r^2 value less than 0.9 with the GTR + F + ASC + R10 maximum-likelihood substitution model (see Methods). d) PCA of global *C. briggsae* isotype reference strains. Colors in each panel represent the sampling geographic region of each strain. The ellipses highlight differentiated isotypes. The numbers in parentheses represent the count of isotypes within each cluster. The colors of the points in panels (b) to (d) correspond to their geographic regions as in panel (a).

KD group), a group from Indonesia (referred to as Indonesia Divergent or ID group), a group Hubei, a group from Quebec, three groups from Taiwan (referred to as Taiwan Divergent 1-3, or TD1-3). A relatedness group of 27 isotypes isolated in locations with temperate latitudes in Asia, Europe, North America, and the Pacific was named after the "Temperate lineage" defined in previous studies (Cutter et al. 2006b; Dolgin et al. 2008; Félix et al. 2013; Thomas et al. 2015). Another relatedness group, named Nairobi–Wisconsin Divergent (NWD), includes two isotypes from Africa (Nairobi) and two isotypes from USA (Wisconsin). Another group of 95 isotypes, which we define as the Taiwan–Hawaii (TH) relatedness group, includes 78 isotypes from Taiwan and 12 isotypes from Hawaii as well as one isotype from each South America, Central America, and the Caribbean and two isotypes from other locations in Asia. The final relatedness group of 501 isotypes found across every sampled geographic region represents the circum-global "Tropical lineage" described in previous studies (Dolgin et al. 2008; Félix et al. 2013; Thomas et al. 2015) (Fig. S7). Overall, these analyses added new isotypes to previously defined relatedness groups (Tropical, Temperate, Hubei, Quebec, Kerala Divergent, and Taiwan Divergent 1) and allowed us to define many newly sampled groups (Australia Divergent, Taiwan Divergent 2 and 3, Indonesia Divergent, and NWD). Despite extensive sampling efforts from the nematode research community, most relatedness groups remain rare and are geographically restricted, with the Tropical group dominating most sampling sites, overshadowed only by the Temperate group in colder climates.

To investigate shared ancestry between relatedness groups, we performed admixture analysis (Alexander et al. 2009). We minimized cross-validation (CV) error by comparing CV estimates for 10 iterations at K assumed subpopulations (from $K = 2$ to 30) and found support for at least 22 subpopulations (see Methods, Figs. S8 and S9). At $K = 22$, nonadmixed individuals from each subpopulation were assigned uniquely to one relatedness group, with certain relatedness groups containing several subpopulations (Fig. S10). Across 10 iterations of $K = 22$, we observed that relatedness groups with low sample sizes were sometimes assigned to the same subpopulation, although relatedness groups with many isotypes remained discrete (Fig. S11). We find admixture among Tropical relatedness group subpopulations as well as admixture of Tropical with subpopulations of the TH relatedness group (Fig. S10), suggesting that these two groups share more recent ancestry, potentially by microhabitat sharing that has enabled cross-breeding. By contrast, most other relatedness groups likely have been evolving independently.

Our results support a history of genetic diversification in Asia and the Pacific, although the geographic origin for *C. briggsae* remains unknown. Previous divergence time estimation dates the oldest split among *C. briggsae* relatedness groups between KD and all other groups (Thomas et al. 2015). The current study uncovered an additional relatedness group from Australia (AD) most closely related to KD but similarly diverged from all other groups, which offers an alternative explanation to a single ancient split in the species evolutionary history (Fig. 2). Because the updated analyses propose a Pacific geographic origin alternative to South India, it is also possible that a third unidentified site proximate to the Pacific and/or Indian Oceans could be the true origin. It is also possible that a relatedness group that could precede the most ancient detected split in this species has not

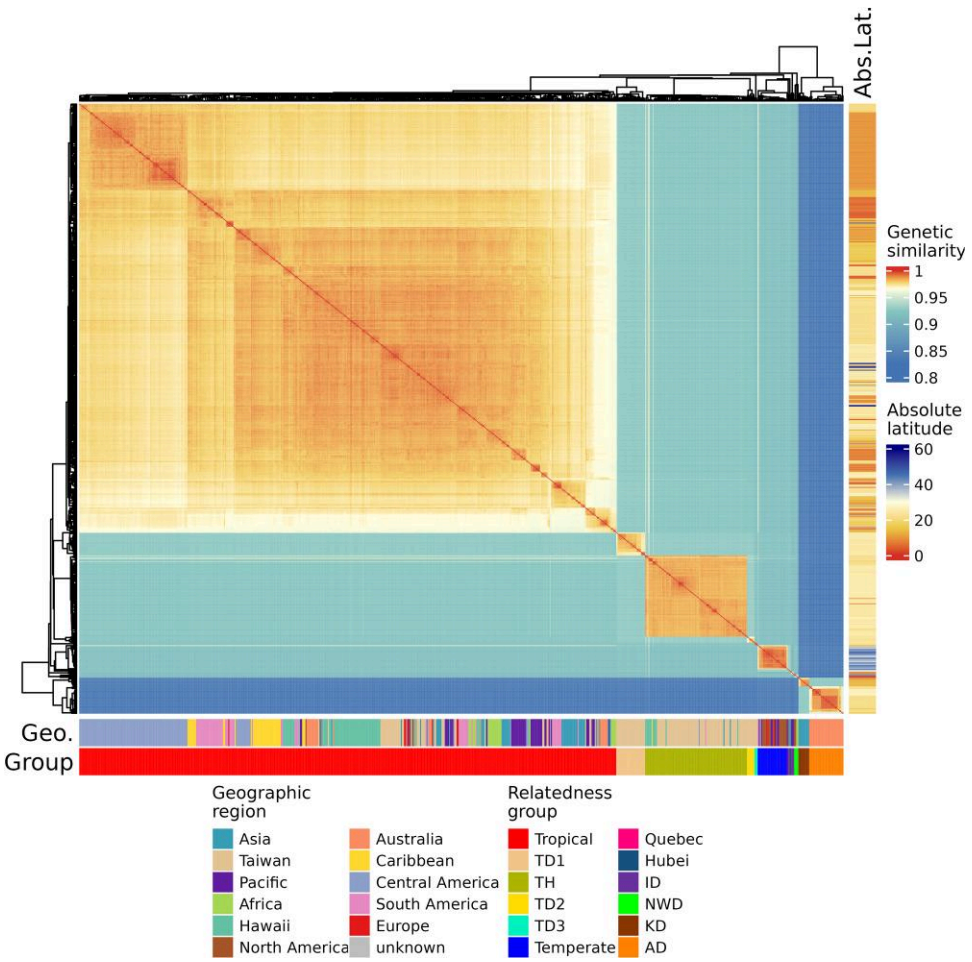

Figure 2 Pairwise genetic similarity defines *C. briggsae* relatedness groups. Heatmap showing pairwise genetic similarity between all 713 isotype reference strains. Pairwise genetic similarity was estimated by the proportion of identical alleles across all identified SNVs among the 713 reference strains. The genetic similarity color gradient has breaks specified at the 25th, 50th, and 75th percentile of the distribution of all pairwise estimates. The right band shows the absolute latitude of the isolation site of each isotype representative strain. The bottom bands show the geographic region of the isolation site of each isotype representative strain and the assigned relatedness group.

been captured among the current sampled strains or is now extinct. The related outbreeding species *C. nigoni* is circum-tropical so the geographic origin of the most recent common ancestor is also unknown (Table S3) and offers no further insights into the geographic origins of *C. briggsae*.

Genetic diversity is concentrated in punctuated HDRs

Population genetics theory predicts self-fertilizing species exhibit lower population-wide genetic diversity than their outbreeding relatives, because frequent self-fertilization increases homozygosity and lowers the effectiveness of recombination (Nordborg and Donnelly 1997; Charlesworth 2012; Cutter 2019). As expected, our global sample of *C. briggsae* exhibits more than 3-fold lower average pairwise nucleotide diversity (π) than *Caenorhabditis remanei* isolated from one location (mean *C. briggsae* $\pi = 0.0047$; mean *C. remanei* $\pi = 0.015$) (Fig. 3, Table S4) (Teterina et al. 2023). Both π and the population mutation rate metric of polymorphism (Watterson's θ (Watterson 1975), θ_W) for *C. briggsae* are higher than those same metrics for *C. elegans* (Table S4) (Crombie et al. 2022). Within *C. briggsae*, we found a 1.9-fold higher mean π and a 1.8-fold higher mean θ_W in arm domains relative to center domains on autosomes (Table S5), consistent with previous analyses of self-fertilizing *Caenorhabditis* nematodes (Thomas et al. 2015; Crombie et al. 2019). This intra-chromosomal difference is attributable to pervasive LD caused by self-fertilization, which intensifies the effects of linked selection in chromosome regions with lower recombination and higher gene density as in the chromosomal center domains (Cutter and Payseur 2013; Cutter 2019). This pattern is less pronounced on the X chromosome, where only a 1.3-fold higher mean π and 1.2-fold higher mean θ_W were measured in arm domains relative to the center domain. At a finer scale, π displays local peaks of nucleotide diversity within relatedness groups (example of Tropical group shown in Fig. S12), which defies the theoretical reduction of genetic diversity caused by self-fertilization. The discovery of HDRs in other self-fertilizing *Caenorhabditis* nematodes, such as *C. elegans* and *C. tropicalis*, provides an attractive explanation for this observation (Lee et al. 2021; Noble et al. 2021).

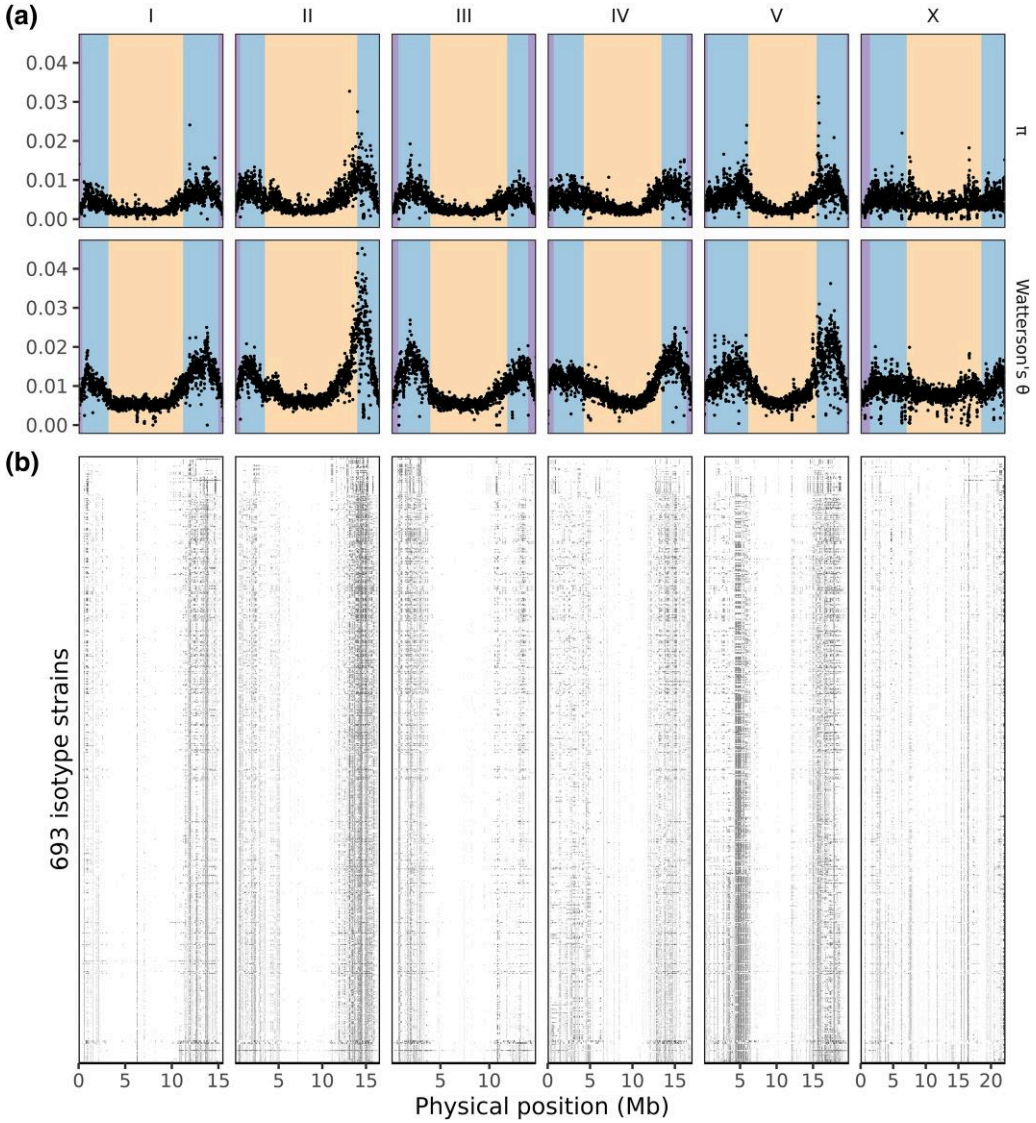

Figure 3 Genetic diversity is concentrated in punctuated HDRs. Each vertical facet separates *C. briggsae* chromosomes with physical coordinates of the QX1410 reference genome on the *x* axis. a) Plots of π and Watterson's θ along genomic positions on each of the six chromosomes. Each point represents a 10 kb window estimate from the pooled sample of all 713 isotype reference strains. Colored backgrounds in nucleotide diversity panels delineate the chromosomal domain boundaries (purple, tip; blue, arm; yellow, center). b) Each row represents one of 693 isotype reference strains, with HDRs marked as black boxes. From a total of 713 isotype reference strains, 20 were excluded because they were used as references for each relatedness group, belonged to relatedness groups that lacked a reference genome assembly, or belonged to relatedness groups that comprised fewer than three isotypes.

To identify and characterize HDRs in *C. briggsae*, we employed methods previously used in *C. elegans* with several modifications (see Methods). In summary, we optimized thresholds to classify and cluster contiguous genomic bins with high variant counts (>11 SNV/kb) or low coverage (<90% bases covered) into HDRs. Given the elevated genome-wide nucleotide diversity between genomes across relatedness groups (Fig. S13), we used multiple reference genomes (Fig. S14), one for each relatedness group, and lifted the genomic coordinates of HDRs onto a single reference coordinate system of the QX1410 genome. Across all relatedness groups, we identified 1,020 nonoverlapping HDRs with an average size of 63 kb, ranging from 5 kb to 3.7 Mb (see Methods; Fig. 3b; Table S6). Collectively, these HDRs covered approximately 60% of the QX1410 reference genome. When assessing relatedness groups individually, we find an average of 763 nonoverlapping HDRs per relatedness group, ranging from a total of 411 (KD group) to 1,361 (Tropical group). Group-specific HDRs range from 5 kb to 1.5 Mb in size and cover from 17% to 38% of the reference genome, comparable to *C. elegans* HDRs (9 kb to 1.35 Mb in size, covering approximately 20% of the N2 genome) (Lee et al. 2021). The Tropical relatedness group has more HDRs compared to all other groups (Fig. S15, Table S7), likely because it comprises the most isotypes (501) from across all sampled geographic regions with a higher likelihood to identify rare hyper-divergent haplotypes.

As expected, HDRs are enriched for SNVs. For example, hyper-divergent haplotypes found in individual isotypes from the Tropical relatedness group contain up to 39% (mean = 31%) of the total alternative alleles across all SNVs, despite covering just 11% or less (mean = 6%) of the reference genome

(Fig. S16, Table S7). Even in the relatedness groups with fewer HDRs than the Tropical group, hyper-divergent haplotypes in individual isotypes cover on average 5% to 12% of their respective relatedness group reference genome and contain on average 37% to 55% of the total alternative alleles across all SNVs called relative to their respective relatedness group reference genomes (Table S7). These results are consistent with hyper-divergent haplotypes in *C. elegans*, where despite covering less than 2% of the genome in individual isotypes they contribute more than 20% of the total alternative alleles across all SNVs on average. The levels of genetic variation across *C. briggsae* HDRs are likely severely underestimated considering that several of these regions can contain novel genes, copy-number variation, gene absences, or genes that are highly divergent and difficult to reliably align to the reference genome (Figs. S17 and S18), much like *C. elegans* (Lee et al. 2021). To quantify the prevalence of gene content variation in HDRs genome wide, we performed enrichment tests for classes of orthologous gene groups with and without variable copy number across 20 isotypes of the Tropical relatedness group (Tables S8 and S9). We found that HDRs contain 34% (2,145/6,282) of orthologous groups with variable copy number, show a significant 1.2-fold enrichment for these variable-copy orthologous groups and are significantly depleted (0.9-fold) for single-copy orthologs (Table S9). The fold enrichment for variable-copy and single-copy orthologs show only minor changes when considering orthologous groups within arm chromosomal domains alone (1.3-fold and 0.88-fold, respectively), suggesting these enrichment results do not originate from a categorical difference between arm and center chromosomal domains (Table S9). By contrast, orthologous groups that contain multiple genes per isotype and the number of genes remain equal across isotypes (namely, "stable-copy" orthologs) show a 1.2-fold enrichment in HDRs when using all orthologous groups but show nonsignificant depletion (0.89-fold) when using orthologous groups in arm chromosomal domains alone (Table S9). These stable-copy orthologs likely represent gene families or paralogous sequences that are overrepresented in arm domains but are not necessarily associated with HDRs.

The highly variable gene content of HDRs offers clues to the possible functional roles for HDRs in *C. briggsae*. Previous gene set enrichment analysis of *C. elegans* HDRs has shown overrepresentation of gene families involved in environmental sensing and pathogen responses, including C-type lectins, G-protein coupled receptors, nuclear hormone receptors, and E3 ubiquitin ligases (eg F-box genes) (Lee et al. 2021). A similar analysis in *C. briggsae* is more difficult because many of the reference genes lack gene ontology (GO) terms. To address this issue, we annotated functional protein domains and their associated GO identifiers using InterProScan (Jones et al. 2014) (see Methods). This analysis identified functional protein domain annotations for approximately 69% (15,289/22,196) of all reference genes (Table S10) and approximately 53% (11,756/22,196) had GO identifiers. Because HDRs are primarily located on chromosome arm domains, we performed an enrichment test for genes in or outside of HDRs in these chromosomal regions (Fig. 4). We find that genes in HDRs on chromosomal arms are significantly enriched for functional protein domains associated with environmental and pathogenic responses, such as G-protein coupled receptors (60/63 genes, class h domain), superfamily domain nuclear hormone receptors (115/137 genes) or orphan *C. elegans*-like nuclear hormone receptors (58/60 genes), and E3 ubiquitin ligases (eg F-box A domains, 68/73 genes; BTB and MATH domains, 52/57 genes; SKP1/BTB/POZ superfamily domains, 85/107 genes) (Fig. 4a). Among enriched GO terms for the same gene sets, we find enrichment for environment sensing behavior such as olfactory behavior (43/61 genes, Fig. 4b) and G protein-coupled peptide receptor activity (47/53 genes, Fig. 4c). We find innate immune response genes like C-type lectins and F-box-containing genes (123/144 genes, eg orthologs of *C. elegans clec-85*, *clec-209*, *clec-218*, *fbxa-141*, and *Y60A3A.8*) (Fig. 4b). Additionally, we find enrichment for xenobiotic metabolism (28/35 genes, Fig. 4b) and monooxygenase activity (39/50 genes, Fig. 4c) involved in xenobiotic metabolism, where many of the genes in both terms were identified as orthologs of *C. elegans* cytochrome P450 families such as *cyp-32*, *cyp-33*, *cyp-34*, and *cyp-35* genes. These results suggest that, like in *C. elegans*, *C. briggsae* HDRs are enriched for environmental response genes that facilitate pathogen response and mediate sensory perception, which likely enables *C. briggsae* to thrive across heterogeneous environments.

Overall, these results reveal that a substantial percentage of variation (on average, 31% to 56% across relatedness groups) is found in punctuated HDRs that cover a relatively small span of *C. briggsae* genomes (on average, up to 11% of individual isotype genomes). Reference genes within HDRs are associated with a multitude of environmental responses, such as olfaction and sensing, transcriptional regulation, and xenobiotic and pathogen stress response. We also find significant enrichment for orthologous gene groups that contain reference genes and exhibit copy-number variation in HDRs relative to the rest of the genome. Additionally, across visualizations of genome assemblies and genes of isotypes within and across relatedness groups, we observe examples of drastic gene content variation in HDRs, including novel genes lacking orthologs in the reference genome. For this reason, a reference-free, pangenome gene set analysis is required to more comprehensively investigate the structural composition and functions of genes found in HDRs that are not represented in the reference genome.

Introgression rarely explains HDRs

Multiple mechanisms have been proposed to explain how HDRs are generated (eg by introgression) and maintained (eg by balancing selection) (Lee et al. 2021; Moya et al. 2025). It is possible that HDRs are not a product of ancestral variation maintained by long-term balancing selection as it has been proposed for *C. elegans* (Lee et al. 2021) but rather represent divergent haplotypes more recently introgressed from its obligate outbreeding sister species, *Caenorhabditis nigoni*. Introgressed divergent haplotypes carrying advantageous alleles could have been preserved as punctuated blocks in *C. briggsae* because of more intense linked selection from self-fertilization. This model of evolutionary rescue in a self-fertilizing population by gene flow from an outcrossing relative is thought to explain highly divergent regions at immunity loci in selfing *Capsella rubella* (Koenig et al. 2019).

To explore the possibility that HDRs could originate from introgression between *C. briggsae* and its outbreeding sister species *C. nigoni*, we needed to develop genomic resources for *C. nigoni*. We sequenced the genomes of eleven *C. nigoni* inbred lines using

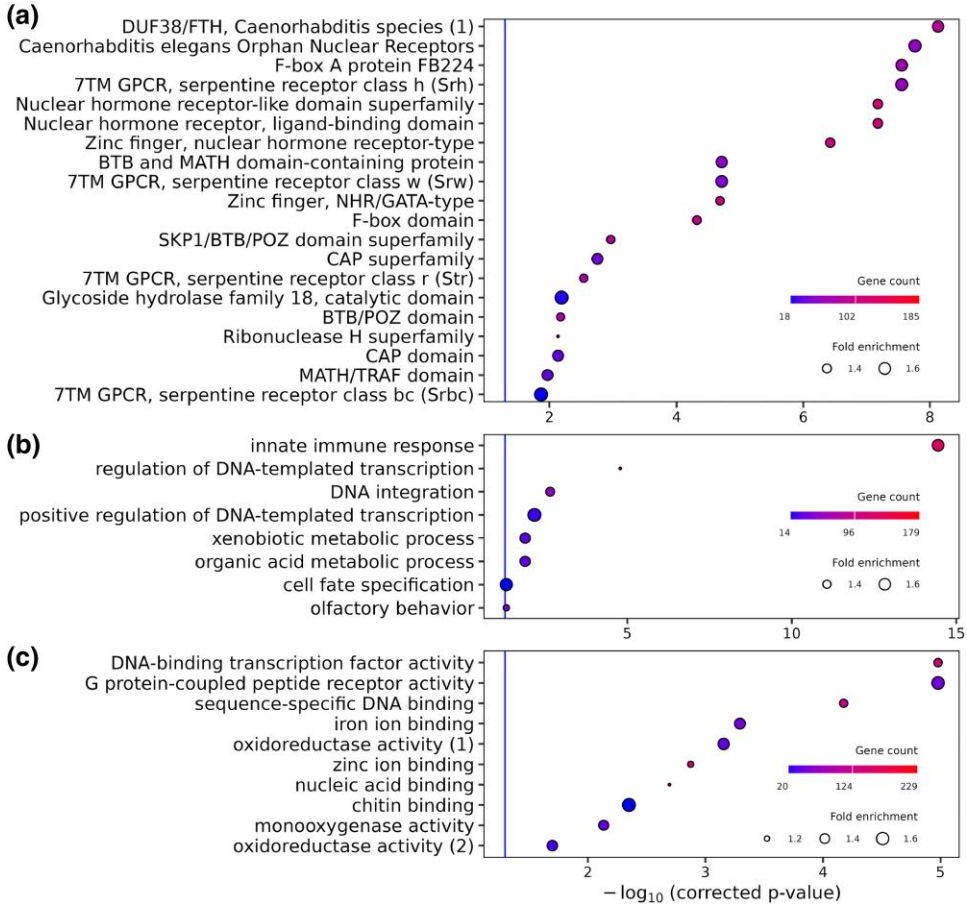

Figure 4 *Caenorhabditis briggsae* HDRs are enriched for environmental response genes. a) Significantly enriched InterProScan functional protein domains in genes found in HDR chromosome arms. Each protein domain annotation on the *y* axis showed a false discovery rate-corrected *P*-value (*q*-value) of lower than 0.05 (negative log-scaled adjusted *P*-values shown on the *x* axis). b) Significant biological process and c) molecular function GO terms from gene set enrichment analysis. Each term present in the *y* axis showed a Benjamini–Hochberg adjusted *P*-value lower than 0.05 (negative log-scaled adjusted *P*-values shown in the *x* axis). In each row, the circle varies in size based on the fold-enrichment associated with that term, and in shading based on the number of genes in chromosome arm HDRs associated with that term. "DUF38/FTH, *Caenorhabditis* species (1)" refers to the gene set term "Domain of unknown function DUF38/FTH, *Caenorhabditis* species", "oxidoreductase activity (1)" refers to the gene set term "oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen", and "oxidoreductase activity (2)" refers to gene set term "oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen, reduced flavin or flavoprotein as one donor, and incorporation of one atom of oxygen". We performed the same tests for chromosomal arm regions that were not classified as hyper-divergent and found no enriched protein domains, and found a few, non-environment related gene ontology terms. All enrichment tests shown use arm domain genes as a background gene set.

PacBio HiFi technology and five additional *C. nigoni* isofemale lines using Oxford Nanopore Technologies (ONT), assembled highly contiguous genomes, and predicted protein-coding genes (see Methods, Table S11). Approaches to detect gene flow under demographic models (eg *fastsimcoal2*) or classical tests for introgression that operate on detecting asymmetry in patterns of allele sharing between taxa (eg Patterson's *D* or *f*-statistics) are difficult because of the limited access to inter-species genome alignment between these *Caenorhabditis* species (Fig. S19), especially in chromosomal arm domains where HDRs are primarily found. This genome accessibility issue is more pronounced when trying to select a potential outgroup species, because the most closely related species with a sufficiently high-quality genome assembly to these sister taxa, *C. remanei*, is separated by 427 million generations of divergence (Fusca et al. 2025) and displays even fewer alignable sequences to *C. briggsae* (Fig. S20).

Instead, we used models of mixture across sites and trees (MAST) (Wong et al. 2024) to assess the distribution of tree topologies in protein sequences across orthologs of *C. briggsae*, *C. nigoni*, and *C. remanei* (Table S8). An excess of sites that reflect a history of introgression or incomplete lineage sorting (ILS) of ancestral variation that has survived since speciation could be indicated by quantitative support for tree topologies discordant from the species tree across protein sequences from those genes found in HDRs, especially juxtaposed to overwhelming support for the species tree among sequences in the rest of the genome. We generated alternative tree topologies for 48 quartets composed of the *C. briggsae* Tropical reference strain (QX1410), one of three potential recipient *C. briggsae* strains from different relatedness groups with approximately 10 Mb of the genome covered by HDRs (JU3237, Tropical; QG4232, AD; JU3200, KD), one of

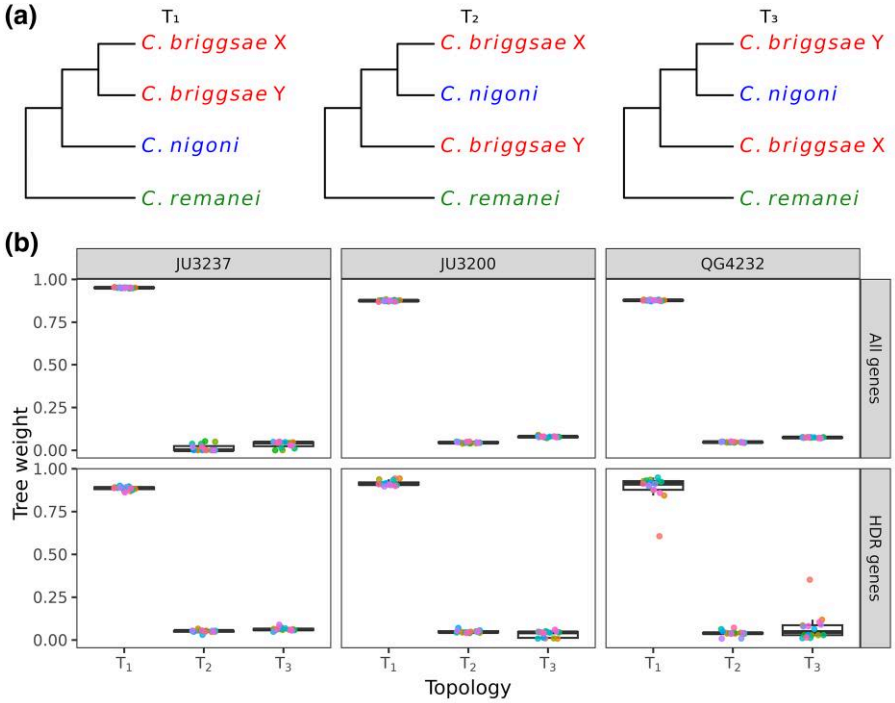

Figure 5 HDRs rarely display an excess of shared alleles with *C. nigoni* relative to the entire genome. a) Three possible tree topologies (T_1, T_2, and T_3) for quartets composed of a reference *C. briggsae* (X) strain from the Tropical group (QX1410), a variable *C. briggsae* (Y) recipient strain from different relatedness groups (JU3237, Tropical; QG4232, AD; JU3200, KD), a variable *C. nigoni* donor (one of 16 inbred lines), and a *C. remanei* outgroup. Topology 1 (T_1) is concordant with the expected species tree, while Topologies 2 and 3 ($T_{2\text{-}3}$) represent discordant topologies where one of the two *C. briggsae* alternative alleles is most closely related to *C. nigoni*. b) Boxplots summarizing the mean and interquantile range of tree weights (*y* axis) assigned to each topology (*x* axis) using MAST (submodel 1) for protein sequence supermatrices assembled for each distinct quartet (shown as points, one color per variable *C. nigoni* donor). Vertical facets separate supermatrices constructed with one the three different *C. briggsae* relatedness group genetic backgrounds (variable recipients), while horizontal facets separate supermatrices constructed exclusively from single-copy orthologs of HDR genes or of all genes in the *C. briggsae* genome.

16 potential donor *C. nigoni* inbred lines, and the outgroup *C. remanei* reference strain (PX506) (Fig. 5a). For each quartet, we used MAST models in IQ-TREE3 to measure the weights assigned to topologies concordant (T_1) or discordant ($T_{2\text{-}3}$) with the expected species tree to assess the frequency of tree mixture across supermatrices of concatenated protein sequences assembled from single-copy orthologous genes (see Methods) found within HDRs compared to estimates of all genes (Fig. 5, Table S12).

Nearly all quartets with variable *C. nigoni* donor lines and *C. briggsae* recipient strains showed no substantial differences in the proportion of weights among alternative topologies between HDRs and the entire genome under MAST submodel 1 (Fig. 5b; Table S12). With the exception of one outlier quartet (QG4232 recipient and ECA2852 donor), T_1 is consistently the most frequent topology (weight greater than 0.84; mean weight = 0.9) and topologies T_2 or T_3 are observed in low frequency (weight less than 0.12; mean weight = 0.05) (Table S12). Repeating this analysis under a more restrictive submodel (submodel 6, see Methods) yielded consistently worse fits according to Bayesian Information Criterion (BIC) (Fig. S21, Table S12) for all quartet supermatrices of all genes (48/48) and for nearly all quartet supermatrices of HDR genes (41/48). When submodel 6 offered a better fit, the weights were congruent with submodel 1 (Table S12). Across quartets in the three variable *C. briggsae* donors, the sum of weights for minor topologies T_2 and T_3 differed by approximately 0.05 between supermatrices of HDR genes and all genes under submodel 1 (Fig. S22). The minor topology weights were relatively higher for HDR supermatrices in quartets that included a strain of the Tropical group (JU3237), relatively lower in quartets that included a strain of the KD group (JU3200), and both lower or higher in quartets that included a strain of the Australia Divergent group (QG4232, varying in direction across *C. nigoni* donors). We also identified asymmetry between the weights of T_2 and T_3 (Fig. S22) that could be consistent with a nonrandom evolutionary process behind the sorting of alleles between *C. briggsae* and *C. nigoni* (analogous to ABBA-BABA test, Green et al. 2010). However, the direction of the asymmetry shifted across quartet compositions (either the weight of T_2 is greater than T_3, or vice versa) and was present in quartet supermatrices constructed from either genes in HDRs or from all genes. Although we failed to detect a clear excess of shared alleles between *C. nigoni* and *C. briggsae* at HDRs relative to the entire genome, the weights assigned to minor topologies were not near zero. Therefore, introgression between the two species could explain diversity in only a minority of sites across the genome. Additionally, we cannot fully discard introgression as an originating mechanism for HDRs, because a different (not sampled and/or extinct) species could have acted as a donor for alleles in HDRs.

C. briggsae HDRs likely arose near the divergence of relatedness groups

The identification of HDRs in *C. elegans* led to the hypothesis that these regions contain adaptive alleles maintained by one or more forms of long-term balancing selection since the evolution of selfing (Lee et al. 2021), likely in selfing nematode populations subject to spatial or temporal variation in environmental pressures. Previous analyses used Tajima's *D* (Tajima 1989) to quantify the allele frequency skews that can suggest different signatures of selection (Lee et al. 2021). We wanted to perform similar analyses of *C. briggsae* HDRs, but the interpretations of Tajima's *D* are biased by the combined influences of demographic size change, direct and linked selection, population structure, and unequal sampling of subpopulations (Städler et al. 2009; Cutter et al. 2012), given the high degree of genetic differentiation between relatedness groups (Fig. 2). To limit these biases, we analyzed Tajima's *D* separately for each relatedness group containing more than ten isotype reference strains (AD, KD, TH, TD1, Temperate, Tropical). Consistent with global *C. briggsae* estimates, we found lower π and θ_W in chromosomal centers relative to arms but to varying degrees (Fig. 6, Figs. S24 and S25, Table S5). Although mean estimates of *D* across chromosomal arms are slightly negative, we find abundant genomic peaks of positive *D* in most relatedness group comparisons (Fig. S26). These peaks are most often found in HDRs (Fig. 6a), suggesting the presence of intermediate-frequency alleles within relatedness groups that could be products of balancing selection. We were curious whether the alleles found in HDRs within relatedness groups significantly diverged between relatedness groups. Estimates of absolute divergence (D_{xy}) in HDRs display higher absolute divergence between relatedness groups relative to non-HDRs (Fig. S27, Table S13). For example, estimates of absolute divergence in arm chromosomal domains between the Tropical and KD relatedness groups is 1.08-fold higher in HDRs, escalating to a maximum fold-increase of 1.47 between the Tropical and TH relatedness groups (Table S13). When compared to D_{xy} at synonymous sites (ssD_{xy}), we find a 1.67-fold increase in the difference of absolute divergence within and outside of HDRs between the Tropical and TH relatedness groups but it is similar between the Tropical and KD relatedness groups (1.13-fold) (Fig. S27, Table S13). The increased levels of neutral variation in HDRs relative to non-HDRs across all relatedness group comparisons are indicative of coalescent times that exceed the split between relatedness groups.

We next sought to place the relative divergence between hyper-divergent haplotypes relative to estimates of divergence between *C. briggsae* and *C. nigoni*. If HDRs have been maintained since speciation as trans-specific polymorphisms, then synonymous site divergence (K_s) between hyper-divergent haplotypes within *C. briggsae* should exceed estimates between the two species. We found that K_s is elevated between hyper-divergent and nonhyper-divergent haplotypes within the Tropical relatedness group (approximately by 0.018; Fig. 6b, Table S14). By contrast, the levels of synonymous site divergence estimated between *C. briggsae* and *C. nigoni* exceed these values by approximately 10-fold ($K_s \sim 0.2$; Fig. 6c, Table S14). This analysis reinforces the conclusions from our gene tree quartet analysis that introgression from *C. nigoni* is unlikely to have introduced HDR haplotypes by introgression between species. K_s between hyper-divergent and nonhyper-divergent haplotypes barely exceeds estimates among nonhyper-divergent haplotypes in comparisons between diverged relatedness groups (Fig. 6b; Table S14). Additionally, in both within- and between-group comparisons, we observed an elevated K_s among hyper-divergent haplotypes relative to among nonhyper-divegent haplotypes (Fig. 6b), especially between comparisons of different relatedness groups, consistent with the presence of multiple alternative haplotypes of a given HDR that have deeper coalescent times than non-HDRs.

A plausible explanation for these observations is that HDRs arose by the establishment of distinct haplotypes followed by isolation (likely both reproductive and geographic) across *C. briggsae* relatedness groups. The initial source of genetic diversity prior to the split of relatedness groups remains elusive, although it is possible that the ancestor of *C. briggsae* remained primarily outcrossing (with ample genetic diversity) for long periods of time prior to the evolution of high self-fertilization. Additionally, it is possible that periods of microhabitat sharing and cross-breeding could have enabled the introduction and establishment of beneficial haplotypes between divergent relatedness groups, but this hypothesis is inconsistent with both elevated estimates of between-group divergence in HDRs presented in this study (Fig. 6; Fig. S27, Table S13) and with the low estimates of natural outcrossing rate of androdioecious *Caenorhabditis* species previously described (Thomas et al. 2015; Gilbert et al. 2022). Instead, the association of HDRs with environmental responses and their observed allele skews toward intermediate frequency suggests that balancing selection could play an important role in the long-term maintenance of hyper-divergent haplotypes within *C. briggsae* relatedness groups. Given that K_s estimates among HDR haplotypes between relatedness groups represent the extreme of neutral divergence in *C. briggsae* (Fig. 6b), balancing selection could be caused by large-scale spatial heterogeneity across environments in distinct geographic ranges and relatedness groups. Within relatedness groups from more narrow geographic ranges, balancing selection caused by temporally varying factors, small-scale spatial environmental differences within microhabitats, or negative frequency dependence with co-occurring pathogen populations might be more plausible. Interestingly, the observation that *C. nigoni* orthologs of genes in *C. briggsae* HDRs also carry increased synonymous site divergence among *C. nigoni* lines (Fig. S28) suggests that both species converge in the distribution of genetic diversity across loci, leaving open the possibility that similar and fluctuating environmental pressures have shaped diversity in both selfing and outcrossing species.

Discussion

The massive expansion of sequenced *C. briggsae* genomes from around the world, enabled by the efforts of the nematode research community, allows for more comprehensive characterization of genomic diversity throughout this species. Our analyses revealed millions of single-nucleotide and thousands of insertion–deletion variants across 713 distinct genome-wide haplotypes (isotypes). These isotypes differ widely in their geographic distribution and relatedness. We discovered regional isotypes comprising large strain cohorts isolated widely across

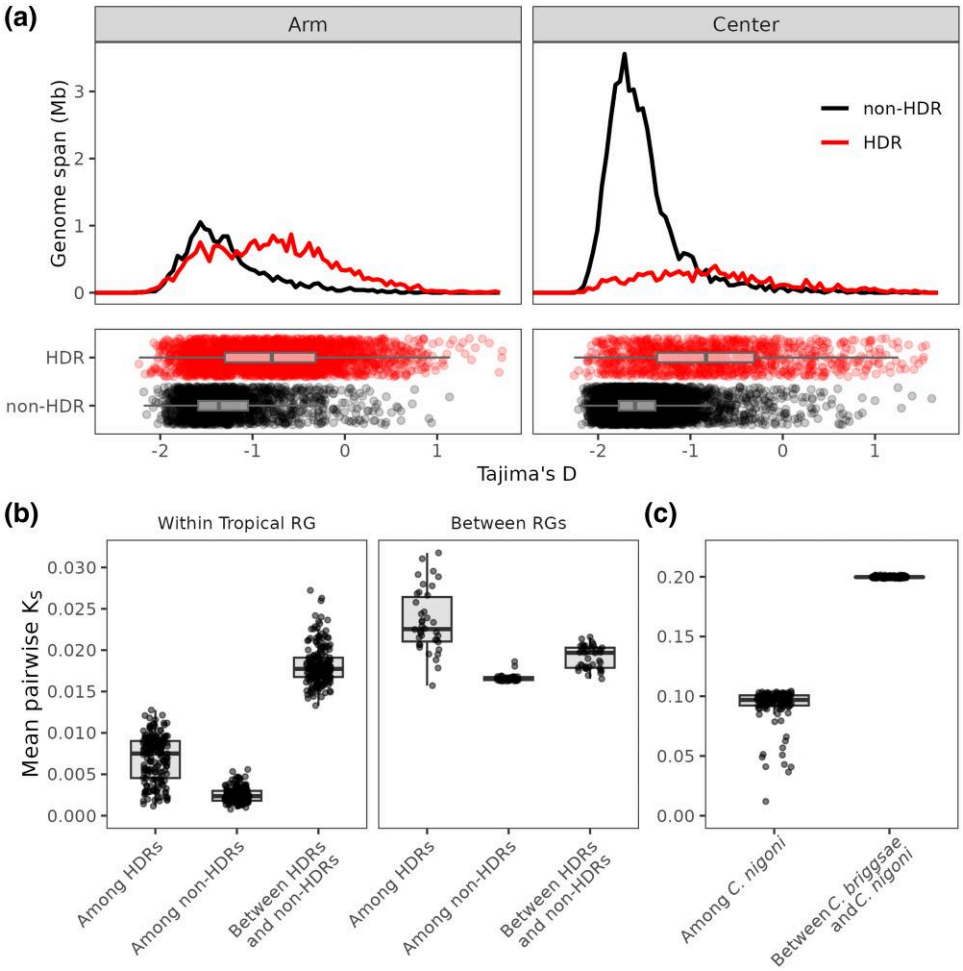

Figure 6 Skew toward intermediate frequency alleles in *C. briggsae* HDRs contrasts lower synonymous site divergence relative to *C. nigoni*. a) Tajima's *D* estimates across 10 kb genome segments among Tropical group isotypes comparing hyper-divergent and non-HDRs. Center and arm estimates are shown as separate vertical facets, with the top panel of each chromosomal domain showing the genome span in the *y* axis against each value of Tajima's *D* on the *x* axis (binned in increments of 0.05), and the bottom panel showing a boxplot comparing hyper-divergent and nonhyper-divergent genome segments. b) Boxplots comparing mean K_s estimates across single copy orthologs between pairs of Tropical relatedness group genomes or between pairs of Tropical and Kerala Divergent or and Tropical and Australia Divergent relatedness group genomes. K_s estimates for each pairwise genome comparison were calculated among hyper-divergent haplotypes, among nonhyper-divergent haplotypes, and between hyper-divergent and nonhyper-divergent haplotypes. c) Boxplots comparing mean K_s estimates across single copy orthologs between pairs of *C. nigoni* genomes or between pairs of *C. nigoni* and (Tropical) *C. briggsae* genomes. In both (b) and (c), each point represents the mean K_s across loci from one genome pair. The boxplot reflects the mean of means (across all genome pairs) and its interquartile range.

continents, underscoring the influence of recent human activity in the dispersal of this species primarily through Europe and North America. Regional isotypes contrast sharply with the majority of isotypes isolated in narrow (<1 km) ranges showing substantial genetic differentiation. We defined 12 relatedness groups that capture the major axes of genetic differentiation, from rare local groups to large globally distributed groups. Previous population genomic studies used highly diverged strains from the south of India (Kerala) to suggest that the earliest diversification of *C. briggsae* could have occurred near the Indian Ocean (Thomas et al. 2015), where its outbreeding sister species *C. nigoni* also is found. The discovery of an Australian relatedness group most closely related to the Kerala relatedness group and similarly diverged from all other groups widens the plausible species origin to the Asia-Pacific region. These results are consistent with the potential geographic origins of several lineages of the Elegans supergroup in the islands of Oceania (Rockman et al. 2026) and emphasize the importance of continued sampling to refine estimates of divergence time and to trace the biogeographic origins of this species.

Our analysis also corroborates chromosomal profiles of genetic diversity observed across self-fertilizing *Caenorhabditis* nematodes such that most genetic variation is concentrated within or near chromosomal arm domains. This uneven distribution of genetic variation along chromosomes reflects the combined effects of linked selection and low recombination in gene-dense central regions. Self-fertilization further exacerbates these effects, producing long tracts of LD and purging genetic variation when selection acts on adaptive or deleterious mutations (Cutter et al. 2006b; Rockman and Kruglyak 2009; Andersen et al. 2012; Thomas et al. 2015; Crombie et al. 2019; Lee et al. 2021; Noble et al. 2021; Stevens et al. 2022). At a finer scale,

nucleotide diversity is condensed in punctuated genomic regions that resemble the recently identified HDRs in *C. elegans* (Lee et al. 2021). Much like *C. elegans*, HDRs in *C. briggsae* encode haplotypes with remarkable gene content variation and encompass large proportions (approximately 30% to 50%) of nucleotide variation within *C. briggsae* relatedness groups. Additionally, *C. briggsae* HDRs also show enrichment for environmental response genes, indicative of the potential role in adaptation for self-fertilizing *Caenorhabditis* species. The characterization of HDRs in *C. briggsae*, accompanied by the contiguous genomes and protein-coding gene models that we have generated for diverse *C. briggsae* wild strains provides new avenues to explore natural variation in this species. We document examples where HDRs implicate novel genes and copy-number variants that yield structurally distinct haplotypes, but it remains to be determined the extent to which different classes of structural variants pervade these regions. Although alignment-based approaches might not be suitable to effectively investigate haplotype diversity in HDRs, the advent of pangenome-based analyses offer a potential solution. Pangenome graphs targeted to HDRs, perhaps associated with phenotypic differences as observed in *C. elegans* (Balla et al. 2015; Greene et al. 2016; Lee et al. 2019; Lee et al. 2023), can help distinguish variation hidden to alignment-based approaches, improve genotype–phenotype associations, and narrow regions of variation to causal genes.

Importantly, HDRs also provide insights into the evolutionary trajectory of *C. briggsae*. Because *C. briggsae* and its sister species *C. nigoni* can hybridize under laboratory conditions (Woodruff et al. 2010; Kozlowska et al. 2012; Bi et al. 2015), we tested whether HDRs in *C. briggsae* might have originated through introgression from *C. nigoni*. We generated high-quality genome assemblies and gene models for *C. nigoni* lines and used our diverse cohort of *C. briggsae* wild strain genomes to identify single-copy orthologs, then used models of MAST to quantify the frequency of shared alleles between *C. briggsae* and *C. nigoni* within HDRs relative to the entire genome. HDRs in *C. briggsae* do not appear to carry an excess of shared alleles with *C. nigoni*, which makes an introgression-based origin for HDRs less likely. This conclusion is reinforced by the magnitude of synonymous site divergence involving HDR haplotypes within *C. briggsae* relative to *C. nigoni*. We inferred the relative age of HDRs to date prior to the split of relatedness groups, which has been estimated to approximately 2 million generations previously (Thomas et al. 2015) but substantially more recent than the divergence of *C. briggsae* with *C. nigoni*. Therefore, we propose that genetic hyper-diversity like that found in other outbreeding *Caenorhabditis* species was once present in the outbreeding ancestor of present-day *C. briggsae* and has since been reduced to punctuated genome segments surrounding hyper-divergent haplotypes.

The emergence of a self-fertilizing lineage leads to rapid reductions in heterozygosity, which exposes recessive deleterious alleles to selection. Selective removal of these recessive deleterious alleles (purging) and linked neutral alleles will lead to further reductions in genetic diversity across a population (Nordborg and Donnelly 1997; Charlesworth 2012; Cutter 2019). However, genetically diverse allelic haplotypes could be retained within a self-fertilizing lineage if they confer sufficiently large context-dependent fitness benefits or occur in genomic regions with a low load of recessive deleterious alleles. Hyper-divergent haplotypes potentially persisted over long periods of time across isolated *C. briggsae* relatedness groups because of varying environmental pressures over space and/or time. It remains unclear whether hyper-divergent haplotypes originated from selection acting on a common ancestral genetic background or from distinct, early intercrossing events between differentiated populations early in *C. briggsae*'s history. HDRs offer an opportunity to study the evolutionary origins of the species, because these regions could be used to approximate the effective population size of the outbreeding ancestor of *C. briggsae*. Such approximations can be used under a simulation framework to set expectations surrounding different models of transition to self-fertilization. Namely, a gradual increase in self-fertilization rate with ample intercrossing with the outbreeding progenitor versus a large-effect self-fertilization origin with limited intercrossing.

We showed that HDR alleles are both distinct among relatedness groups and present at intermediate-frequency alleles within *C. briggsae* relatedness groups. These patterns are consistent with genomic regions under balancing selection at both the species scale, potentially reflecting ecological adaptation by distinct relatedness groups, and at the subpopulation scale, such that different haplotypes are adaptive under temporally varying environmental factors or spatially varying microhabitat conditions encountered by members of a given relatedness group. Although our findings do not resolve the mode of balancing selection that could have enabled the persistence of such genetic diversity, the kinds of overrepresented genes encoded within HDRs suggest some hypotheses. HDRs in selfing *Caenorhabditis* species including *C. briggsae* are enriched for gene classes suggesting diverse environmental responses, including pathogen, xenobiotic stress, and sensory responses. Divergent pathogen resistance alleles in HDRs could persist under negative frequency-dependent selection imposed by pressure from diverse pathogen populations. Similar evolutionary dynamics are thought to contribute toward remarkable allelic diversity maintained at vertebrate major histocompatibility complex loci (Takahata and Nei 1990). Additionally, divergent alleles at pathogen, olfaction, and xenobiotic stress response loci could be maintained under spatially varying selection from different types of bacteria or fungi at diverse sampled locations or temporally varying within the same location. To distinguish among these selective processes, an analysis that integrates distinct hyper-divergent haplotypes and ecological data is necessary. Ecological sampling of microbial communities where *Caenorhabditis* nematodes have been isolated could help quantify exposure to natural pathogens (eg *Pseudomonas* spp. or *Nematocida parisii*) and help establish associations with locations where distinct hyper-divergent haplotypes were found. Near-isogenic lines carrying distinct hyper-divergent haplotypes or allele replacements of candidate causal loci could be tested for fitness advantages against natural pathogens found in a particular locality and provide evidence for the influence of heterogeneous environments in preserving divergent haplotypes in a local populations of *Caenorhabditis* nematodes.

Within the *Caenorhabditis* genus, several species now have worldwide samples comprising thousands of wild isolates, whole-genome sequences, and a wide array of chromosomal or near-chromosomal assemblies. Additionally, *Caenorhabditis* nematodes are highly tractable experimental systems because

of their short generation time, validated toolkits for CRISPR genome editing, and robust platforms for high-throughput phenotyping. Overall, these newly created strain resources for *C. briggsae* and those already available for related species enable comparative genomics studies currently not possible in other animal species. Genome-wide association studies in *C. briggsae* can be performed using specific relatedness groups (eg Tropical) to minimize the effects of population structure, thereby reducing false discovery of quantitative trait loci (QTL) (Andersen and Rockman 2022). By comparing quantitative variation in traits shared between *C. briggsae* and *C. elegans*, the field could identify functionally conserved loci not possible by studying a single species in isolation. These approaches are now possible and future high-throughput assays that take advantage of large strain sets from both species should be studied to make significant impacts on our understanding of evolutionary processes.

The *C. briggsae* wild strains and their genomic data are all deposited in the *Caenorhabditis* Natural Diversity Resource, establishing an unprecedented platform for comparative genomics across multiple species from the *Caenorhabditis* genus (Cook et al. 2017; Crombie et al. 2024). This resource enables a variety of molecular quantitative genetics approaches from QTL discovery to the validation of causal variants found in quantitative trait genes (Gaertner and Phillips 2010; Evans et al. 2021; Andersen and Rockman 2022). *Caenorhabditis briggsae* strain diversity paired with high-throughput assays to measure quantitative traits will give the community a large collection of QTL to answer questions related to population and quantitative genetics. Allele frequencies and predicted variants can be parsed from these QTL and compared to QTL (and orthologous genes) from *C. elegans* (Evans et al. 2021). Likewise, analyses of variation in orthologous genes between *C. briggsae* and *C. elegans* and across other *Caenorhabditis* species facilitate studies of repeated or convergent evolution when combined with robust quantitative phenotyping. We believe that the incorporation of *C. briggsae* species-wide variation can broadly impact evolutionary genetics beyond the larger *Caenorhabditis* community.

Methods

Strains

A total of 1,972 *C. briggsae* strains and 16 *C. nigoni* inbred lines were included in this study (Tables S1 and S3). *Caenorhabditis briggsae* strains and 11 *C. nigoni* inbred lines (ECA2852, ECA2857, JU1422, JU1418, JU1419, JU1420, NIC2143, NIC2150, NIC2152, VX151, VX153) were reared at 20 °C using *Escherichia coli* bacteria (strain OP50) grown on modified nematode growth medium (NGMA) (Andersen et al. 2014) containing 1% agar and 0.7% agarose to prevent animals from burrowing. All *C. briggsae* isotype reference strains used in this study have been deposited at and are available upon request from the *Caenorhabditis* Natural Diversity Resource (CaeNDR) (Crombie et al. 2024). The remaining five *C. nigoni* isofemale lines (EG5268, JU2484, JU2617, YR106, ZF1220) for ONT sequencing were cultured at 25 °C on NGM plates with twice the standard agar concentration to prevent nematode burrowing. The following divergent strains were checked for cross compatibility with AF16 using a fluorescently labeled strain (QG2801): QG4064, QG4094, QG4095, QG4131, QG4208, QG4233, JU2801, JU3202, JU3203, and JU3206. All strains gave rise to viable progeny.

Whole-genome sequencing

DNA extraction was performed following established protocols (Cook et al. 2016). In brief, to extract DNA, we transferred nematodes from three 10 cm NGMA plates spotted with OP50 *E. coli* into a 15 mL conical tube by washing with 10 mL of M9. We then used gravity to settle animals on the bottom of the conical tube, removed the supernatant, and added 10 mL of fresh M9. We repeated this wash method three times 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 DNAEasy Blood and Tissue isolation kit cat# 69506 (cat# 69506, QIAGEN, Valencia, CA). The DNA concentration was determined for each sample with the Qubit dsDNA Broad Range Assay Kit cat# Q32850 (cat# Q32850 Invitrogen, Carlsbad, CA). All strains were used to construct short-read libraries using NEBNext Ultra II FS DNA Library Prep (cat# E6177L). These libraries were sequenced on several Illumina NovaSeq platforms (paired-end 150 bp reads) at the Duke Center for Genomic and Computational Biology, Novogene, NUSeq, the New York University core facility, and the Johns Hopkins Genetics Resources Core Facility. The raw short-read sequencing data are available from the NCBI Sequence Read Archive (Project PRJNA1149159).

We sequenced the genomes of 59 *C. briggsae* strains using PacBio HiFi sequencing. The selected *C. briggsae* strains represent a manually selected panel of strains from diverse source laboratories, geographic regions, and variable levels of divergence from the QX1410 reference genome, with divergence approximated from branch lengths from a strictly bifurcating phylogeny built from SNV sites across all isotypes sequenced using short-read technology (see *Alignments and variant calling* and Species-wide tree construction sections below). Out of the 59 *C. briggsae* samples, 13 (BRC20341, VX34, NIC1660, ECA2670, QG4132, ECA2666, BRC20492, JU3237, ECA2617, QX1796, TWN1824, NIC1529, JU3200) were each grown on 15 10-cm plates spotted with OP50 *E. coli*. The nematodes were washed off of the plates with M9 and allowed to settle by gravity. The pellets were washed three times using M9 and then frozen at −80 °C. All of these pellets were processed (DNA extraction and library generation) and sequenced at the Cold Spring Harbor Laboratory Genome Center. For the remaining 46 *C. briggsae* (Table S9), and 11 *C. nigoni* samples (ECA2852, ECA2857, JU1422, JU1418, JU1419, JU1420, NIC2143, NIC2150, NIC2152, VX151, VX153), animals were grown on four to five 10 cm NGMA plates spotted with OP50 *E. coli*. Just prior to starvation, the animals were washed off with cold M9 into a 50 mL conical and spun at 2,000 rpm for 8 min with a break of three. The supernatant was removed and the pellet was washed three times with cold M9 + 0.01% Tween. A final wash was performed with cold M9. The majority of the supernatant was removed and the pellet was transferred to a 1.7 mL microfuge tube using a glass Pasteur pipette. The animals were spun at 2,000 rpm for 3 min, the supernatant was removed, the pellet was flash frozen, and then stored at −80 °C. The pellet should weigh no more than 60 mg. To make HMW DNA, we used the Monarch HMW DNA Extraction Kit for Tissue (cat# T3060L,

NEB, Ipswich, MA). The protocol was completed with the following modifications/specifics: Prior to using the pestle, we added 50 µL of the lysis buffer to the sample and added the remaining 600 µL after homogenization; for the lysis at 56 °C after homogenization, we incubated in a Thermomixer for 15 min at 750 rpm and then at 550 rpm for 30 min or longer; use cold Protein Separation Solution; leave on ice for 10 min after the inversion step; use three Capture beads; to elute, we added 200 µL Elution Buffer, incubated for 5 min at 56 °C and 300 rpm, another 50 µL of Elution Buffer was added prior to separating the DNA from the beads; the final solubilization was done at 37 °C for 30 min at 300 rpm. Library generation and sequencing were carried out by Maryland Genomics at the University of Maryland or Duke Center for Genomic and Computational Biology. For ONT sequencing of five *C. nigoni* isofemale lines (EG5268, JU2484, JU2617, YR106, ZF1220), *C. nigoni* genomic DNA of each strain was extracted from at least six starved nematodes on 90 mm plates using the MasterPure Complete DNA and RNA Purification Kit (Biosearch Technologies). R10.4.1 flow cells (FLO-MIN114) and the ligation sequencing kit (SQK-LSK114) were then used to further process genomic DNA with 4 sµg as the input for each sequencing run. Sequencing and basecalling were carried out on a GridION platform using MinKNOW software v5.9.18 with default parameters, and only high-quality reads were kept for genome assembly.

Alignments and variant calling

Adapters and low-quality sequences were removed from raw Illumina reads and trimmed reads shorter than 20 bp were discarded using *fastp* v0.20.0 and default parameters (Chen 2023). Subsequently, *C. briggsae* QX1410 (project: PRJNA784955) (Stevens et al. 2022) was used as reference genome on which to align Illumina reads with Burrows–Wheeler Aligner, *BWA-MEM* v0.7.17-r1188 (Li 2013). *Sambamba* v0.7.0 (Tarasov et al. 2015) was used to merge and index libraries of the same strain, where duplicate sequencing reads were flagged with *Picard* v2.21.3 (https://broadinstitute.github.io/picard/). Strain coverage was calculated using mosdepth v0.2.6 (Pedersen and Quinlan 2018). Strains with less than 10× coverage were excluded from subsequent analyses.

Variant calling for each strain was conducted using *HaplotypeCaller* in GATK v4.1.4.0 (Poplin et al. 2018). Then, the identified variants were aggregated and jointly recalled using *GenomicsDBImport* and *GenotypeGVCFs* in *GATK* v4.1.4.0 (Poplin et al. 2018). The variants were further processed for polarizing heterozygous SNVs and quality filtering with the Andersen Lab pipeline *wi-gatk* (https://github.com/AndersenLab/wi-gatk). Biallelic heterozygous SNVs were adjusted to homozygous reference (REF) or alternate (ALT) if there was substantial read support for this conversion. In this step, we only looked at SNVs in nucleic DNA. Specifically, the heterozygous SNV was converted if the normalized Phred-scaled likelihoods (PL) met the following criteria, where smaller PL values indicate higher confidence: SNVs were converted to homozygous ALT if PL-ALT/PL-REF $\leq$ 0.5 and PL-ALT $\leq$ 200; likewise, they were converted to homozygous REF if PL-REF/PL-ALT $\leq$ 0.5 and PL-REF $\leq$ 200. Any heterozygous SNVs that failed to meet these criteria were not modified in this step. At the site-level, variants were subjected to several filters: ensuring variant quality (QUAL) > 30; normalizing variant quality by read depth (QD) > 20; mitigating strand bias of ALT calls with strand odds ratio (SOR) < 5 and fisherstrand (FS) < 100; Limiting the fraction of samples with missing genotype to <95%; and restricting fraction of samples with heterozygous genotype after heterozygous site polarization to <10%. At the sample level, filtration of variants were based on read depth (DP) > 5. Any remaining heterozygous sites were converted to missing (./.), except for heterozygous sites on mitochondria, which were not changed. After these procedures, invariant sites were removed. The generated population-wide variant call format (VCF) file was referred to as the "hard-filtered VCF file".

In addition, a second VCF file was produced to include all invariant site calls for population statistics. The genotyping was performed as described above with the exception that at the joint-genotyping step, the "–all-sites" flag was used to prevent invariant sites from being dropped. The resulting joint variant calls were then split into variant and invariant sites based on having an allele count greater than one or less than or equal to one, respectively. All variant sites were processed as described above. Invariant sites were processed as above with the following exceptions. Invariant sites were not subjected to any allele frequency filtering, heterozygous polarization, or filtering based on QUAL, QD, SOR, or FS, as these flags are not present for invariant sites. All sites were then concatenated and sorted by chromosome and position to produce a VCF file containing genotype calls for all reference sites referred to as the "all-sites VCF file".

Isotype reference strain characterization

A total of 1,972 strains were grouped into 713 highly related groups called isotypes using pairwise genetic similarity. First, genetic similarity was estimated from the ratio of identical alleles relative to the total nonmissing SNV sites (total SNVs: 5,756,001) between each pair of strains computed using a custom script with the hard-filtered VCF file as the input. Second, strains were then hierarchical agglomerative clustered by genetic similarity using complete linkage with a cutoff of 99.97% similarity, such that all pairwise similarities between strains within a given isotype group exceeded this cutoff. The threshold of 99.97% genetic similarity was selected based on the pairwise genetic similarity distribution, which aims to remove redundant (nearly identical) genotypes from our analyses (Fig. S29). This method is available as a Nextflow pipeline (https://github.com/andersenlab/isotype-nf). Genetic similarity estimates across all strain pairs or across all isotype reference strain pairs are available on GitHub (see Data availability).

Divergence time approximation in the NIC174 regional isotype

Fourfold degenerate sites were annotated with the QX1410 reference genome and protein-coding gene annotations using *degenotate* v1.3 (https://github.com/harvardinformatics/degenotate). We then subsetted the hard-filtered VCF file to 4-fold degenerate sites between the two least genetically similar strains in the NIC174 isotype (JU2220 and MY754). Assuming a *C. elegans* mutation rate (μ) of 2.3×10^{-09} per site per generation

(Saxena et al. 2019) and using the proportion of identified 4-fold degenerate site variants ($k = 30/6{,}690{,}038$) between the two least genetically similar strains in the NIC174 isotype, we applied a molecular clock under an infinite sites model (Jukes and Cantor 1969; Kimura 1985) to approximate a divergence time ($T \approx k/2\mu$) of 975 generations.

Relatedness of isotype reference strains

Relatedness of the *C. briggsae* isotype reference strains were analyzed based on the LD-pruned VCF file, which was generated from the hard-filtered VCF file after the following variants-pruning steps: Genome-wide biallelic SNPs on all chromosomes were extracted and normalized by joining separate biallelic variants into multiallelic records using *bcftools norm -m +* in bcftools v1.14. Then the biallelic-only SNPs were extracted and LD pruned with PLINK v1.90b6.21 (with *–indep-pairwise 50 10 0.9*) to remove highly correlated markers (Purcell et al. 2007; Chang et al. 2015). Finally, after removal of missing sites from the markers, the loci corresponding to the remaining markers were extracted from the normalized VCF file to generate an LD-pruned VCF containing only biallelic variants without missing sites. Hereafter, this file is referred to as the "LD-pruned VCF file". PCA was performed with smartpca v16000 executable from EIGENSOFT v7.2 (Patterson et al. 2006; Price et al. 2006). The LD-pruned marker files (.ped, .snp, and.ind) produced by PLINK were supplied as input to smartpca to calculate Eigenstrat values and Tracy–Widom statistics. A total of 219 PCs were required to explain 80% of the genetic variance, and 144 PCs contributed substantially to explaining genetic variation in that they exhibited eigenvalues >1.

Species-wide tree construction

We used the *vcf2phylip.py* script (Ortiz 2019) to convert the LD-pruned VCF files into the PHYLIP (Felsenstein 1993) format. Then, IQ-TREE2 v2.4.0 (Minh et al. 2020) was used to construct a tree with PHYLIP format file as input. The GTR + F + ASC + R10 substitution model was selected as the maximum-likelihood model under automatic model selection via ModelFinder based on the BIC (Kalyaanamoorthy et al. 2017). We limited our model search to extensions of the GTR substitution model (*-mset GTR).* Node support was assessed using 1,000 ultra-fast bootstrap (UFBoot2) replicates (*–bb 1000*). Trees were then visualized in R v4.3.2 with the ggtree v3.10.1 R package (Yu et al. 2017). Pairwise genetic similarity among 713 global *C. briggsae* isotype reference strains were clustered and visualized with the ComplexHeatmap v2.18.0 R package (Gu 2022).

Assignment of the regional isotype reference strains

All the isotypes that contain any strains with pairwise geographic distances over 1 km among sampling locations were defined as regional isotypes. The geographic distances were calculated with the *geosphere* v1.5-18 R package. The Haversine formula was used to calculate the distance, which assumed the shape of Earth is a perfect sphere (Karney 2013).

Repeat masking

To mask the genomic repeats, we identified *C. briggase* genome-wide repetitive elements using methods described previously (Stevens et al. 2022) with minor modifications. First, repetitive sequences were identified modeled de novo using RepeatModeler from RepeatMasker v2.0.6 (Smit et al. 2015). Transposable elements were then identified using TransposonPSI v1.0.0 (http://transposonpsi.sourceforge.net). Long terminal repeat (LTR) retrotransposons were identified with LTRharvest and then annotated with LTRdigest from GenomeTools v1.6.5 (Ellinghaus et al. 2008; Gremme et al. 2013), with HMM profiles from the Gypsy Database v2.0 (Llorens et al. 2011) and Pfam v37.4 domains (Mistry et al. 2021). Repetitive sequences were then further filtered with the gt-select tool from GenomeTools to remove sequences lacking conserved protein domains. We then obtained all Rhabditid repeats from RepBase (Bao et al. 2015) and Dfam (Hubley et al. 2016). The generated repeat libraries were merged into a single redundant repeat library, clustered with VSEARCH v2.22.1 (Rognes et al. 2016) from QIIME2 v2024.10 (Bolyen et al. 2019), and then classified with the RepeatClassifier tool from RepeatModeler. Finally, the unclassified repeats that had BLASTX v2.16.0, (Camacho et al. 2009) hits (custom parameters: *-evalue 0.001*) to *C. elegans* or *C. briggsae* proteins were removed using RepeatMasker to eliminate genuine gene sequences that had been misidentified as repeats.

Population genomic statistics

Genome-wide nucleotide diversity (π), Watterson's θ (θ_W), Tajima's D, and absolute divergence (D_{xy}) were calculated with *pixy* v2.0.0 (Korunes and Samuk 2021; Bailey et al. 2025) for 10 kb windows using the all-sites VCF file as the input, with relatedness groups and isotype reference strains defined under the *–populations* argument. Synonymous-site D_{xy} were also estimated using *pixy* and the same arguments and input files with 4-fold degenerate site positions included in the *–sites* argument. All reported estimates associated with a specific chromosomal domain (arms, center, or tip) used breakpoints previously defined in the QX1410 genome (Stevens et al. 2022). These chromosomal domain breakpoints were defined by steep changes in recombination rate in a recombination map constructed from SNV markers genotyped in 99 recombinant inbred lines generated from crosses between the QX1410 and VX34 strains.

Admixture analysis

The admixture analysis was performed using ADMIXTURE v1.3.0 (Alexander et al. 2009) with the PED file as the input. This PED file contains the LD-pruned sites and was exported by PLINK when generating the LD-pruned VCF file. Ten independent times of 10-fold CV were used to determine the optimized number of total populations where the curve of CV error first reached a minimization plateau. The total number of tested populations ranged from 2 to 30. Nonadmixed representative strains were defined by the maximum subpopulation fraction over 99.9%. CV error was minimized by comparing CV estimates across 10 runs for each

assumed K value (from $K = 2$ to 30) and found support for at least 22 subpopulations (Fig. S8). At $K = 22$, nonadmixed individuals from each subpopulation were assigned uniquely to one relatedness group, with certain relatedness groups containing several subpopulations (Fig. S10). Across 10 independent runs under $K = 22$, five relatedness groups were sometimes assigned to the same subpopulation, although the remaining relatedness groups consistently stayed discrete (Fig. S11).

Long-read genome assembly and protein-coding gene prediction

Sequenced PacBio HiFi genomic reads were demultiplexed and adapters were removed by the sequencing cores using using *lima* (Duke: v2.9.0, CSHL: v2.2.0, UMD: v2.12.0; https://lima.how). Demultiplexed and adapter-trimmed samples from the same strain sequenced from different facilities were merged using *samtools* v1.20 (Danecek et al. 2021). PCR duplicates were removed using *pbmarkdup* v1.1.1 (https://github.com/PacificBiosciences/pbmarkdup). Primary assemblies were generated from deduplicated HiFi FASTAs using *hifiasm* v0.24.0-r702 (Cheng et al. 2021) with the initial bloom filtered disabled (*-f0*, recommended for small genomes) and haplotig purging disabled (*-l0*, recommended for homozygous genomes). Genome assemblies generated from merged samples from the same strain were compared against assemblies generated from each individual sample, and the assembly with the highest N50 was kept. After assembly, contigs were taxonomically classified using DIAMOND 2.1.13 (Buchfink et al. 2021) with parameters *–faster* and an *e*-value of 1E−10 against the UniProt database (Release 2025_03) (UniProt Consortium 2025). Contigs that were annotated as non-*Nematoda* were purged using BlobToolKit v4.4.5 (Challis et al. 2020) with *–param bestsumorder_phylum–Inv = Nematoda*. For assembly of *C. nigoni* ONT genomic reads, raw reads were first corrected using Canu v2.2 (Koren et al. 2017) with the parameter *-correct genomeSize=130 m*. The corrected reads were assembled with Flye v2.9.3 (Kolmogorov et al. 2019) under the *-nano-raw* setting optimized for nanopore data. Misassemblies in the contigs were then corrected using the RagTag *–correct* module, which was executed with the *ragtag.py* (Alonge et al. 2022) correct command and MUMmer v4.0.0 (Marçais et al. 2018) as the aligner. The alignment was performed with the *nucmer* tool from the MUMmer package using the parameters *–mum –mincluster 100 –maxgap 300*. Subsequently, the *–scaffold* module in RagTag was used to anchor the corrected contigs to the *C. nigoni* reference genome JU1421. To remove contaminants, bacterial scaffolds were removed based on BLAST v2.11.0 searches against the NCBI nt database with an *e*-value threshold of 1E−9. The remaining scaffolds were then underwent five rounds of polishing with Pilon v1.24 (Walker et al. 2014) by mapping Illumina short reads to the assemblies using BWA v2.2.1 to improve sequence accuracy. For each assembled genome, protein-coding gene models were predicted with BRAKER3 v3.0.8 (Gabriel et al. 2023) trained with a custom library of protein sequences from *C. elegans* N2 (WormBase, WS283) and *C. briggsae* QX1410 (PRJNA784955) for *C. briggsae* assemblies or protein sequences from Eukaryota OrthoDB v12 (Kuznetsov et al. 2023) for *C. nigoni* assemblies. Genome and gene model biological completeness was estimated using BUSCO v5.0.4 (Simão et al. 2015) with the *nematoda_odb10* database.

Selection of relatedness group reference genomes for group-specific HDR calling

Detection of HDRs in *C. elegans* relied on clustering one kilobase (kb) partitions of the genome that exceeded empirically determined thresholds of SNVs count and relative coverage depth using short-read sequencing alignments between wild strains and the laboratory strain (N2) reference genome (Lee et al. 2021). Using a single reference genome to call HDRs species-wide in *C. elegans* was likely only possible because of minimal population structure driven by large-scale selective sweeps (Andersen et al. 2012). Considering previous reports of comparisons in *C. briggsae* between strains from the Tropical and Kerala relatedness groups that displayed a lack of marked differences in absolute sequence divergence among chromosomal domains (Thomas et al. 2015), and similar observations in the distribution of genome-wide SNV between divergent groups (Fig. S13), we expected that the increased density of single-nucleotide variation expected in HDRs would be masked by standing genetic variation stemming from divergence between relatedness groups. To overcome this limitation, we called HDRs in each *C. briggsae* relatedness group using group-specific reference genomes. Relatedness groups composed of less than three isotypes or that lacked a reference genome were excluded. A total of seven relatedness group reference genomes were selected (TD2, BRC20492; TD1, BRC20530; AD, ECA2670; KD, JU1348; Temperate, JU2536; TH, NIC1660; Tropical, QX1410) (Fig. S14).

Identification of HDRs

A schematic of the HDR calling workflow described below is available (Fig. S30). HDR calling was carried out using a custom R script available on GitHub (see Data availability). To classify genomic regions as hyper-divergent, we first aligned 15 long-read genome assemblies from strains in the Tropical relatedness group (ECA1657, ECA176, EG6267, JU3207, JU3237, NIC1529, QG1005, QG2641, QG2665, QG2892, QG2902, QG2964, QG789, QG791, and QX1796) against the Tropical reference genome (QX1410) using *nucmer* v3.1 (Marçais et al. 2018) (custom parameters, –maxgap 500 –mincluster 100) (Kurtz et al. 2004). Alignments smaller than 1 kb were discarded. The QX1410 reference genome was then partitioned into 1 kb bins using *bedtools* v2.31.1 *makewindows* (Quinlan and Hall 2010). In each strain, the alignment identity of every bin was estimated from the average identity of the alignment covering the bin. When multiple alignments of variable length and identity covered a bin, the longest alignment was used to estimate the bin identity. The percentage of bases covered by at least one alignment in each bin was also estimated. Bins under 96% identity or 60% bases covered were classified as hyper-divergent (Fig. S31). To further refine region boundaries, nonhyper-divergent bins that were flanked by hyper-divergent bins on both sides were re-classified as hyper-divergent (gap re-classification step). Contiguous hyper-divergent bins were clustered into HDRs (bin clustering step).

Next, we filtered the hard-filtered VCF containing variant calls from all 713 isotype reference strains down to the 500 isotype reference strains in the Tropical relatedness group using *bcftools* v1.21 (Danecek et al. 2021). Only the variant sites where at least one of the 500 isotype reference strains in the Tropical relatedness group had an ALT allele were kept after subsetting. For each strain, we estimated the SNV count in each reference bin using *bcftools* v1.21 (Danecek et al. 2021) and *bedtools coverage* v2.31.1 (Quinlan and Hall 2010), and the percentage of bases covered in every reference bin using *mosdepth* v0.3.10 (Pedersen and Quinlan 2018). We tested a wide range of threshold pairs, including raw variant count (5 to 30 SNVs per kb, 1 SNV step) paired with percent bases covered thresholds (5% to 90%, 5% step), to classify and cluster bins into HDRs using short-read alignment data for the 15 strains in the Tropical relatedness group where long-read-based HDRs were previously called. We selected the optimal variant count and percent bases covered thresholds for short-read HDR calls by concordance between short- and long-read HDR call boundaries. We estimated the concordance of short- and long-read HDR call boundaries by measuring the overlap fraction (proportion of a long-read HDRs that overlapped a short-read HDRs) and the excess fraction (proportion of a short-read call that exceeded the boundary of an overlapping long-read call).

For each strain, we estimated the mean overlap fraction and mean excess fraction across every overlap between the short- and long-read HDR calls at every threshold pair (Figs. S32 and S33). We also estimated the completeness and accuracy of the short-read hyper-divergent calls by calculating recall (proportion of long-read calls that had any overlap with short-read calls), and precision (proportion of short-read calls that had any overlap with long-read calls) (Figs. S34 and S35). Recall and precision were used to estimate an *F*1 score for each threshold pair (Fig. S36). We selected the top *N* (starting at $N = 1$) threshold pairs based on the *F*1 score for each strain, and increased the value of *N* by 1 until we identified a threshold pair that was present among all of the seven strains in this comparison (referred to as "consensus optimal"). The consensus optimal threshold pair was reached at $N = 10$ (top 2.2% of all threshold pairs), where the optimal thresholds selected were a variant count of 11 paired with 90% bases covered (Fig. S37). We classified bins with estimated variant count or percent bases covered under these optimal thresholds as hyper-divergent, genome-wide, across all 500 isotype reference strains in the Tropical relatedness group. We then performed the same gap re-classification, bin clustering steps described for long-read based HDRs. HDRs separated by less than 5 kb were merged (region clustering step), and HDRs smaller than 5 kb were discarded (size filtering step). For each of the remaining relatedness groups, we scaffolded each reference genome relative to QX1410 using *ragtag.py* v2.1.0 (Alonge et al. 2022) (custom parameters, –mm2-params "-x asm20") and aligned short-read DNA sequences from each isotype reference strain against their respective relatedness group reference genome, called variants using GATK (see Alignments and variant calling section above), and called HDRs using the same methods and optimal thresholds described for the Tropical relatedness group. We mapped the positions of HDRs called in non-Tropical reference genomes to the QX1410 reference genome using genome alignments generated with *nucmer* v3.1 (Marçais et al. 2018) (Fig. S38). When genome alignments did not span one or both boundaries of a HDR, we extended the region to the most proximal alignment (up to 50 kb distance) outside of the region boundary or dropped the region when no proximal alignment was found.

Functional enrichment analysis

InterProScan v5.75 (Jones et al. 2014) was used to identify functional protein domains and their associated GO IDs that are enriched in genes found in HDR arm domains and non-HDR arm domains. InterProScan was run with options –goterms, –iprlookup, and –disable-precalc on the longest-isoform proteome of QX1410 created using *agat_sp_keep_longest_isoform.pl* from AGAT v1.4.1 (Dainat et al. 2024) and *gffread* v0.12.7 (Pertea and Pertea 2020). A total of 15,289 QX1410 proteins were annotated with InterProScan functional protein domains, and 11,756 proteins were annotated with GO IDs. Of the 15,289 proteins with domain annotations, genes found on chromosome arms were used as the background set for enrichment analysis, resulting in 5,301 genes being used. Of the 11,756 proteins with GO ID annotations, 3,905 genes found on chromosome arms were used as the background set for GO enrichment analysis. Of the reference genes found within HDRs on chromosomal arms, approximately 62% (3,097/4,956) received InterProScan functional protein domain annotations and approximately 43% (2,1114/4,956) received GO identifiers. All QX1410 genes found on chromosome arms were separated based on their presence or absence in HDRs identified among all 500 *C. briggsae* isotype reference strains belonging to the Tropical relatedness group. For InterProScan domain enrichment analysis, a one-sided hypergeometric test was conducted in R, with a false discovery rate adjusted *P*-value using the R package *stats* v.4.2.1. InterProScan functional protein domains were classified as significantly enriched if they had an adjusted *P*-value less than 0.05. GO enrichment analysis was performed using the R package clusterProfiler v4.6.2 (Yu et al. 2012; Wu et al. 2021; Xu et al. 2024; Yu 2024). GO terms were classified as significantly enriched if they met the criteria of having both a Benjamini–Hochberg adjusted *P*-value and a *q*-value of less than 0.05. To identify QX1410 orthologs of N2, OrthoFinder v3.1.0 (Emms et al. 2025) was run using proteomes translated from the reference protein-coding gene annotations of QX1410 (Moya et al. 2023) and N2 (WormBase, WS283) using only the longest isoforms for each protein-coding gene model.

Identification of orthologous genes

We filtered for the longest isoform across protein-coding gene annotations of 16 *C. nigoni* lines, 20 Tropical *C. briggsae* strains, 2 *C. briggsae* strains from divergent relatedness groups (QG4232, AD group; JU3200, KD group), and reference protein sequences of an outgroup *C. remanei* strain (PX506) using *agat_sp_keep_longest_isoform.pl* from AGAT v1.4.1 (Dainat et al. 2024), followed by protein sequence extraction using *gffread* v0.12.7 (Pertea and Pertea 2020). Orthologous groups across the 39 taxa were identified with OrthoFinder v3.1.0 (Emms et al. 2025) using custom parameters: *-S diamond_ultra_sens -A mafft*. Orthogroup and gene IDs can be accessed in the supplement material

(Table S8). We identified a total of 27,584 orthologous groups. From all orthologous groups, 15,062 shared between at least one member from the three species and 1,768, 5,136, and 595 were unique to either *C. briggsae*, *C. nigoni*, or *C. remanei*, respectively. From the shared orthologous groups, 4,564 represent single-copy orthologs from which 3,977 were present across all *C. briggsae*, *C. nigoni*, and *C. remanei* individuals.

Quartet design for tree topology analysis

A total of 48 distinct quartets were designed using four-taxon arrays with variable *C. nigoni* donor and *C. briggsae* recipient strains. The reference Tropical *C. briggsae* strain (QX1410) and the reference *C. remanei* outgroup strain (PX506) were included across all quartets. One of three alternative *C. briggsae* potential recipient strains (JU3237, Tropical; QG4232, AD; JU3200, KD) were paired with one of 16 potential *C. nigoni* donor lines (see *Whole-genome sequencing* for list of *C. nigoni* lines) across the 48 quartets. We generated three alternative topologies ($T_{1\text{-}3}$) for each quartet, where T_1 represents the expected species tree (both *C. briggsae* taxa are closely related) and $T_{2\text{-}3}$ represent discordant trees where one of two *C. briggsae* taxa are most closely related to *C. nigoni* than to one another (Fig. 5a).

Assembly of quartet amino-acid supermatrices

Amino-acid supermatrices were assembled from protein sequences in orthologous groups using PhyloPyPruner v1.3.0 (https://gitlab.com/fethalen/phylopypruner). Alignments between sequences of each orthologrous group were constructed with MAFFT v7.525 (Katoh and Standley 2013). Alignments were supplied to PhyloPyPruner paired with a consensus amino-acid substitution tree modeled with IQ-TREE v3.0.1 (Wong et al. 2026) using the "LG" substitution model (*-m LG*) and 1,000 replicates of ultra-fast bootstrap *(-bb 1,000)*. Quartet composition across supermatrices was handled by passing a list of samples not belonging to the focal quartet to the *–exclude* argument. Pre-filtering arguments to collapse nodes with less than 75% support into polytomies (*–min-support 0.75*), remove sequences shorter than 20 amino acids (*–min-len 20*), remove loci with less four taxa (*–min-taxa* 4), remove loci with less than 80% sequence occupancy (*–min-gene-occupancy 0.8*), and remove sequences with a branch length that is seven times larger than the standard deviation of all branches within its tree (*–trim-lb 7*, effectively excluding outlier protein sequences from inaccurate gene models) were kept constant across all assembled supermatrices. The argument to exclude all sequences from taxa with less than 10% matrix occupancy (*–min-otu-occupancy 0.1*) was also enabled, although no taxa were removed across all assembled supermatrices. Selection of single-copy orthologs for each quartet supermatrix was handled with the *–trim* argument (*–trim 1to1*). An alternative supermatrix consisting of only single-copy orthologs within HDRs was generated for each focal quartet by excluding orthologous groups where the reference QX1410 sequence had less than a 50% overlap with an HDR called in the alternative recipient *C. briggsae* strain of the focal quartet.

MAST analysis

Each quartet supermatrix, composed of either all single-copy orthologs or single-copy orthologs within HDRs, was passed to IQ-TREE v3.0.1 (Wong et al. 2026) along with a MAST submodel (Wong et al. 2024) and a list of pre-defined topologies ($T_{1\text{-}3}$, see the section "Quartet design for tree topology analysis"). We tested the least restricted submodel (submodel 1, unlinked AA frequencies, substitution rate, and RHAS) (Fig. 5b) and most restrictive submodel (submodel 6, linked AA frequencies, substitution rate, and RHAS) (Fig. S21).

Synonymous site divergence

For each sample, protein-coding DNA sequences were retrieved using *gffread* v0.12.7 (Pertea and Pertea 2020). Protein sequence alignments of each orthologous group from OrthoFinder (see Identification of orthologous genes) were reverse translated into codon-aware coding sequence alignments using *PAL2NAL* v14 (Suyama et al. 2006). Synonymous site divergence for each alignment was estimated with the *kaks()* function of the *seqinr* R package (https://github.com/lbbe-software/seqinr) using the default estimator (Li 1993) and default gap handling (rmgap = TRUE, codons with gaps in any sample are removed). Divergence estimates within and between species were processed separately by subsetting the alignments to sequences pertinent to each comparison within R (eg removing sequences from *C. nigoni* when making divergence estimates within *C. briggsae*). R scripts are available on GitHub (see Data availability).

Acknowledgments

We thank members of the Andersen Lab for providing comments on this manuscript. We especially thank Loraina A. Stinson, Claire Buchanan, Kristen M. Laricchia, Shannon C. Brady, Lauren Seeley, Kreena Patel, Scott E. Baird, Nausicaa Poullet, Julien Dumont, Deborah Bourc'his, Romain Salle, Gunther Hollopeter, Hinrich Schulenburg, Carola Peterson, Yen-Ping Hsueh, Ching-Ting Yang, Chia-Yi Kao, Jessica N. Sowa, Jay Ni, Carolynn M. O'Donnell, Fabien Duveau, Amhed Vargas Velazquez, Isabelle Nuez, Jean-Baptiste Pénigault, Gautier Brésard, Alexandre Peluffo, Fabrice Besnard, Henrique Teotonio, Amir Yassin, Christopher Nelson, V. Robert, H. Baylis, A. Akay, M. Volovitch, R. Ferrière, M. Murmu, Renaud de Rosa, L. Lokmane, M. Joron, V. Debat, N. Gidaszewski, L. Barre, A. Evin, C. Yujin Kim, D. Peigne, I. Felix, Bradford Heights Elementary School and Francis S. Grandinetti Elementary School Nematode Hunters, and citizen scientists for contributing wild *C. briggsae* strains to CaeNDR. We would also like to thank Michael G. Tassia and Lewis Stevens for their guidance in analyses of introgression and sequence divergence.

Author contributions

N.D.M. and E.C.A. conceived and designed the study. N.D.M., B.W., L.M.O., M.E.G.S., and E.C.A. analyzed the data. N.D.M., B.W., L.M.O., R.E.T., M.E.G.S., and E.C.A. wrote the manuscript.

R.E.T., M.E.G.S., A.O.S., R.M., M.A., A.B.P., Z.Z., A.D.C., M.V.R., and M.-A.F. edited the manuscript. R.E.T., A.K, and N.M.R. performed whole-genome sequencing for 1,972 *C. briggsae* wild isolates. R.E.T. and A.K. performed long-read sequencing for 59 *C. briggsae* wild isolates and 11 *C. nigoni* inbred lines. R.E.T., C.G., C.M.D., T.A.C., G.Z., E.R., L.F., V.D.D., E.H., M.D., C.G., D.E.C, J.C.H., A.S., S.Z., A.R., T.W., A.M., S.H., K.R.A., E.J.K., N.M.R., E.S.S., V.S., H.T., M.A., A.B.P., Z.Z., A.D.C., J.W., M.V.R., M.-A.F., C.B., and E.C.A. performed sample collection and isolation of 1,972 *C. briggsae* strains. M.E.G.S. performed isotype characterization analysis for 1,972 *C. briggsae* wild strains.

Supplementary material

Supplementary material is available at *Molecular Biology and Evolution* online.

Funding

This work was funded by National Science Foundation capacity grant 2224885, Human Frontiers Science Program grant RGP0001/2019, and the National Institutes of Healthgrant R01ES029930 (to E.C.A). C.B. acknowledges support by the Centre National de la Recherche Scientifique (CNRS), the Institut National de la Santé et de la Recherche Médicale (Inserm), and Université Côte d'Azur. This work was also funded by the following agencies: Centre National de la Recherche Scientifique, and a grant of the Indo-French Centre for the Promotion of Advanced Research Ref. 6503-4 (to V.S. and M.A.F.); NIHNational Institute of General Medical Sciences grant R35 GM141906 (to M.V.R.); Biodiversity Research Center (Academia Sinica, Taiwan) internal funds, National Science and Technology Council (Taiwan) grant NSC 100-2311-B-001-015-MY3, and Academia Sinica grant 103-CDA-L01 (to J.W.); National Science Foundation CAREER grant MCB-1552101 (to M.A.); Hong Kong Environment Conservation Fund, ECF-160/2023 and General Research Grants (GRF), 12101522, 12100024, and 12101323 (to Z.Z.); NSF Building Research Capacity of New Faculty in Biology (BRC-BIO) grants 2218079 (to Jessica N. Sowa); National Institutes of Health NIEHS grant R50ES037948 (to M.E.G.S.); and National Institutes of Health grant F32AI181342 (to A.O.S.).

Conflicts of interest

None declared.

Data availability

The raw short-read sequencing reads for the wild *C. briggsae* strains are publicly available from the NCBI Sequence Read Archive (PRJNA1149159). The raw PacBio long-read data are publicly available from the NCBI Sequence Read Archive (PRJNA1406752). Strain information and short-read genomic variation data are available from the CaeNDR (www.caendr.org) (Crombie et al. 2024). Scripts and related processed data are available at GitHub (https://github.com/AndersenLab/CB_PopGen_HDR_MS/).

References

Abrams MB *et al.* Population and comparative genetics of thermotolerance divergence between yeast species. *G3 (Bethesda)*. 2021:11:jkab139. https://doi.org/10.1093/g3journal/jkab139.

Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res*. 2009:19:1655–1664. https://doi.org/10.1101/gr.094052.109.

Alonge M *et al.* Automated assembly scaffolding using RagTag elevates a new tomato system for high-throughput genome editing. *Genome Biol*. 2022:23:258. https://doi.org/10.1186/s13059-022-02823-7.

Andersen EC, Rockman MV. Natural genetic variation as a tool for discovery in Caenorhabditis nematodes. *Genetics*. 2022:220:iyab156. https://doi.org/10.1093/genetics/iyab156.

Andersen EC *et al.* Chromosome-scale selective sweeps shape *Caenorhabditis elegans* genomic diversity. *Nat Genet*. 2012:44:285–290. https://doi.org/10.1038/ng.1050.

Andersen EC, Bloom JS, Gerke JP, Kruglyak L. A variant in the neuropeptide receptor npr-1 is a major determinant of *Caenorhabditis elegans* growth and physiology. *PLoS Genet*. 2014:10:e1004156. https://doi.org/10.1371/journal.pgen.1004156.

Bailey N, Stevison L, Samuk K. Correcting for bias in estimates of θw$ þ_w and Tajima's D D $ from missing data in next-generation sequencing. *Mol Ecol Resour*. 2025:25:e14104. https://doi.org/10.1111/1755-0998.14104.

Baird SE, Chamberlin HM. *Caenorhabditis briggsae methods*. WormBook; 2006. p. 1–9. https://doi.org/10.1895/wormbook.1.128.1.

Balla KM, Andersen EC, Kruglyak L, Troemel ER. A wild *C. elegans* strain has enhanced epithelial immunity to a natural microsporidian parasite. *PLoS Pathog*. 2015:11:e1004583. https://doi.org/10.1371/journal.ppat.1004583.

Bao W, Kojima KK, Kohany O. Repbase update, a database of repetitive elements in eukaryotic genomes. *Mob DNA*. 2015:6:11. https://doi.org/10.1186/s13100-015-0041-9.

Bi Y *et al.* A genome-wide hybrid incompatibility landscape between *Caenorhabditis briggsae* and *C. nigoni*. *PLoS Genet*. 2015:11:e1004993. https://doi.org/10.1371/journal.pgen.1004993.

Bolyen E *et al.* Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nat Biotechnol*. 2019:37:852–857. https://doi.org/10.1038/s41587-019-0209-9.

Buchfink B, Reuter K, Drost H-G. Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nat Methods*. 2021:18:366–368. https://doi.org/10.1038/s41592-021-01101-x.

Camacho C *et al.* BLAST+: architecture and applications. *BMC Bioinformatics*. 2009:10:421. https://doi.org/10.1186/1471-2105-10-421.

Cao J *et al.* Whole-genome sequencing of multiple Arabidopsis thaliana populations. *Nat Genet*. 2011:43:956–963. https://doi.org/10.1038/ng.911.

Challis R, Richards E, Rajan J, Cochrane G, Blaxter M. BlobToolKit—interactive quality assessment of genome assemblies. *G3 (Bethesda, Md.)*. 2020:10:1361–1374. https://doi.org/10.1534/g3.119.400908.

Chang CC *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience*. 2015:4:7. https://doi.org/10.1186/s13742-015-0047-8.

Charlesworth B. The effects of deleterious mutations on evolution at linked sites. *Genetics*. 2012:190:5–22. https://doi.org/10.1534/genetics.111.134288.

Chen S. Ultrafast one-pass FASTQ data preprocessing, quality control, and deduplication using fastp. *Imeta*. 2023:2:e107. https://doi.org/10.1002/imt2.107.

Cheng H, Concepcion GT, Feng X, Zhang H, Li H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat Methods*. 2021:18:170–175. https://doi.org/10.1038/s41592-020-01056-5.

Coghlan A, Stajich JE, Harris TW. Comparative genomics in *C. elegans*, *C. briggsae*, and other *Caenorhabditis* species. *Methods Mol Biol*. 2006:351:13–29. https://doi.org/10.1385/1-59745-151-7:13.

Cole R, Holroyd N, Tracey A, Berriman M, Viney M. The parasitic nematode *Strongyloides ratti* exists predominantly as populations of long-lived asexual lineages. *Nat Commun*. 2023:14:6427. https://doi.org/10.1038/s41467-023-42250-1.

Cook DE *et al.* The genetic basis of natural variation in *Caenorhabditis elegans* telomere length. *Genetics*. 2016:204:371–383. https://doi.org/10.1534/genetics.116.191148.

Cook DE, Zdraljevic S, Roberts JP, Andersen EC. CeNDR, the *Caenorhabditis elegans* natural diversity resource. *Nucleic Acids Res*. 2017:45:D650–D657. https://doi.org/10.1093/nar/gkw893.

Crombie TA *et al.* Local adaptation and spatiotemporal patterns of genetic diversity revealed by repeated sampling of *Caenorhabditis elegans* across the Hawaiian Islands. *Mol Ecol*. 2022:31:2327–2347. https://doi.org/10.1111/mec.16400.

Crombie TA *et al.* Deep sampling of Hawaiian *Caenorhabditis elegans* reveals high genetic diversity and admixture with global populations. *eLife*. 2019:8:e50465. https://doi.org/10.7554/eLife.50465.

Crombie TA *et al.* CaeNDR, the caenorhabditis natural diversity resource. *Nucleic Acids Res*. 2024:52:D850–D858. https://doi.org/10.1093/nar/gkad887.

Cutter AD. Reproductive transitions in plants and animals: selfing syndrome, sexual selection and speciation. *New Phytol*. 2019:224:1080–1094. https://doi.org/10.1111/nph.16075.

Cutter AD, Payseur BA. Genomic signatures of selection at linked sites: unifying the disparity among species. *Nat Rev Genet*. 2013:14:262–274. https://doi.org/10.1038/nrg3425.

Cutter AD, Baird SE, Charlesworth D. High nucleotide polymorphism and rapid decay of linkage disequilibrium in wild populations of *Caenorhabditis remanei*. *Genetics*. 2006a:174:901–913. https://doi.org/10.1534/genetics.106.061879.

Cutter AD, Félix M-A, Barrière A, Charlesworth D. Patterns of nucleotide polymorphism distinguish temperate and tropical wild isolates of *Caenorhabditis briggsae*. *Genetics*. 2006b:173:2021–2031. https://doi.org/10.1534/genetics.106.058651.

Cutter AD, Wasmuth JD, Washington NL. Patterns of molecular evolution in Caenorhabditis preclude ancient origins of selfing. *Genetics*. 2008:178:2093–2104. https://doi.org/10.1534/genetics.107.085787.

Cutter AD, Wang G-X, Ai H, Peng Y. Influence of finite-sites mutation, population subdivision and sampling schemes on patterns of nucleotide polymorphism for species with molecular hyperdiversity: FINITE-SITES MUTATION, POPULATION SUBDIVISION AND SAMPLING SCHEMES. *Mol Ecol*. 2012:21:1345–1359. https://doi.org/10.1111/j.1365-294X.2012.05475.x.

Dainat J, *et al. NBISweden/AGAT: AGAT-v1.4.1*. Zenodo, 2024. https://zenodo.org/records/13799920

Danecek P *et al.* Twelve years of SAMtools and BCFtools. *Gigascience*. 2021:10:giab008. https://doi.org/10.1093/gigascience/giab008.

Dolgin ES, Félix M-A, Cutter AD. Hakuna Nematoda: genetic and phenotypic diversity in African isolates of *Caenorhabditis elegans* and *C. briggsae*. *Heredity (Edinb).* 2008:100:304–315. https://doi.org/10.1038/sj.hdy.6801079.

Ellinghaus D, Kurtz S, Willhoeft U. LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. *BMC Bioinformatics*. 2008:9:18. https://doi.org/10.1186/1471-2105-9-18.

Emms DM, Liu Y, Belcher LJ, Holmes J, Kelly S. 2025 July 16. OrthoFinder: scalable phylogenetic orthology inference for comparative genomics [preprint]. bioRxiv 2025.07.15.664860. https://www.biorxiv.org/content/10.1101/2025.07.15.664860v1.abstract

Evans KS, van Wijk MH, McGrath PT, Andersen EC, Sterken MG. From QTL to gene: C. elegans facilitates discoveries of the genetic mechanisms underlying natural variation. *Trends Genet*. 2021:37:933–947. https://doi.org/10.1016/j.tig.2021.06.005.

Eves-van den Akker S *et al.* The genome of the yellow potato cyst nematode, *Globodera rostochiensis*, reveals insights into the basis of parasitism and virulence. *Genome Biol*. 2016:17:124. https://doi.org/10.1186/s13059-016-0985-1.

Félix M-A, Duveau F. Population dynamics and habitat sharing of natural populations of *Caenorhabditis elegans* and *C. briggsae*. *BMC Biol*. 2012:10:59. https://doi.org/10.1186/1741-7007-10-59.

Félix M-A *et al.* Species richness, distribution and genetic diversity of Caenorhabditis nematodes in a remote tropical rainforest. *BMC Evol Biol.* 2013:13:10. https://doi.org/10.1186/1471-2148-13-10.

Felsenstein J. *PHYLIP (Phylogeny inference package), version 3.5 c*. Joseph Felsenstein, 1993. http://www.dbbm.fiocruz.br/molbiol/main.html

Fusca DD, Dall'Acqua MN, Sánchez-Ramírez S, Cutter AD. Phylogenomic timetree-calibrated speciation clocks for Caenorhabditis nematodes reveal slow but disproportionate accumulation of post-zygotic reproductive isolation. *PLoS Genet*. 2025:21:e1011852. https://doi.org/10.1371/journal.pgen.1011852.

Gabriel L *et al.* 2023 Feb 29. BRAKER3: Fully automated genome annotation using RNA-seq and protein evidence with GeneMark-ETP, AUGUSTUS and TSEBRA [preprint]. bioRxiv 2023.06.10.544449. https://pmc.ncbi.nlm.nih.gov/articles/PMC10312602/

Gaertner BE, Phillips PC. *Caenorhabditis elegans* as a platform for molecular quantitative genetics and the systems biology of natural variation. *Genet Res (Camb).* 2010:92:331–348. https://doi.org/10.1017/S0016672310000601.

Gilbert KJ *et al.* The distribution of mutational effects on fitness in *Caenorhabditis elegans* inferred from standing genetic variation. *Genetics*. 2022:220:iyab166. https://doi.org/10.1093/genetics/iyab166.

Graustein A, Gaspar JM, Walters JR, Palopoli MF. Levels of DNA polymorphism vary with mating system in the nematode

genus caenorhabditis. *Genetics*. 2002:161:99–107. https://doi.org/10.1093/genetics/161.1.99.

Green RE *et al.* A draft sequence of the Neandertal genome. *Science*. 2010:328:710–722. https://doi.org/10.1126/science.1188021.

Greene JS, Dobosiewicz M, Butcher RA, McGrath PT, Bargmann CI. Regulatory changes in two chemoreceptor genes contribute to a *Caenorhabditis elegans* QTL for foraging behavior. *eLife*. 2016:5:e21454. https://doi.org/10.7554/eLife.21454.

Gremme G, Steinbiss S, Kurtz S. GenomeTools: a comprehensive software library for efficient processing of structured genome annotations. *IEEE/ACM Trans Comput Biol Bioinform*. 2013:10:645–656. https://doi.org/10.1109/TCBB.2013.68.

Gu Z. Complex heatmap visualization. *iMeta*. 2022:1:E43. https://doi.org/10.1002/imt2.43.

Hillier LW *et al.* Comparison of *C. elegans* and *C. briggsae* genome sequences reveals extensive conservation of chromosome organization and synteny. *PLoS Biol*. 2007:5:e167. https://doi.org/10.1371/journal.pbio.0050167.

Hubley R *et al.* The Dfam database of repetitive DNA families. *Nucleic Acids Res*. 2016:44:D81–D89. https://doi.org/10.1093/nar/gkv1272.

Jones P *et al.* InterProScan 5: genome-scale protein function classification. *Bioinformatics*. 2014:30:1236–1240. https://doi.org/10.1093/bioinformatics/btu031.

Jovelin R, Ajie BC, Phillips PC. Molecular evolution and quantitative variation for chemosensory behaviour in the nematode genus Caenorhabditis. *Mol Ecol*. 2003:12:1325–1337. https://doi.org/10.1046/j.1365-294X.2003.01805.x.

Jukes TH, Cantor CR. Evolution of protein molecules. In: *Mammalian protein metabolism*. Elsevier; 1969, p. 21–132. https://doi.org/10.1016/B978-1-4832-3211-9.50009-7.

Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat Methods*. 2017:14:587–589. https://doi.org/10.1038/nmeth.4285.

Karney CFF. Algorithms for geodesics. *J Geodesy*. 2013:87:43–55. https://doi.org/10.1007/s00190-012-0578-z.

Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol*. 2013:30:772–780. https://doi.org/10.1093/molbev/mst010.

Kimura M. 1985. The neutral theory of molecular evolution. https://books.google.com/books?hl=en&lr=&id=e_HoAwAAQBAJ&oi=fnd&pg=PR9&ots=TsxVxpI-9w&sig=T6YvaIcxIEvQCGB1mOnaPqlLohw

Kiontke KC *et al.* A phylogeny and molecular barcodes for Caenorhabditis, with numerous new species from rotting fruits. *BMC Evol Biol.* 2011:11:339. https://doi.org/10.1186/1471-2148-11-339.

Klein J. Generation of diversity at MHC loci: implications for T-cell receptor repertoires. *Immunology*. 1980:80:239–253.

Koenig D *et al.* Long-term balancing selection drives evolution of immunity genes in Capsella. *eLife*. 2019:8:e43606. https://doi.org/10.7554/eLife.43606.

Kolmogorov M, Yuan J, Lin Y, Pevzner PA. Assembly of long, error-prone reads using repeat graphs. *Nat Biotechnol*. 2019:37:540–546. https://doi.org/10.1038/s41587-019-0072-8.

Koren S *et al.* Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res*. 2017:27:722–736. https://doi.org/10.1101/gr.215087.116.

Korunes KL, Samuk K. pixy: unbiased estimation of nucleotide diversity and divergence in the presence of missing data. *Mol Ecol Resour.* 2021:21:1359–1368. https://doi.org/10.1111/1755-0998.13326.

Kozlowska JL, Ahmad AR, Jahesh E, Cutter AD. Genetic variation for postzygotic reproductive isolation between *Caenorhabditis briggsae* and *Caenorhabditis* sp. 9: genetic variation for postzygotic reproductive isolation. *Evolution*. 2012:66:1180–1195. https://doi.org/10.1111/j.1558-5646.2011.01514.x.

Kurtz S *et al.* Versatile and open software for comparing large genomes. *Genome Biol*. 2004:5:R12. https://doi.org/10.1186/gb-2004-5-2-r12.

Kusaba M *et al.* Self-incompatibility in the genus Arabidopsis: characterization of the S locus in the outcrossing *A. lyrata* and its autogamous relative *A. thaliana*. *Plant Cell*. 2001:13:627–643. https://doi.org/10.1105/tpc.13.3.627.

Kuznetsov D *et al.* OrthoDB v11: annotation of orthologs in the widest sampling of organismal diversity. *Nucleic Acids Res*. 2023:51:D445–D451. https://doi.org/10.1093/nar/gkac998.

Kyes SA, Kraemer SM, Smith JD. Antigenic variation in *Plasmodium falciparum*: gene organization and regulation of the var multigene family. *Eukaryot Cell*. 2007:6:1511–1520. https://doi.org/10.1128/EC.00173-07.

Lee D *et al.* Selection and gene flow shape niche-associated variation in pheromone response. *Nat Ecol Evol*. 2019:3:1455–1463. https://doi.org/10.1038/s41559-019-0982-3.

Lee D *et al.* Balancing selection maintains hyper-divergent haplotypes in *Caenorhabditis elegans*. *Nat Ecol Evol*. 2021:5:794–807. https://doi.org/10.1038/s41559-021-01435-x.

Lee D *et al.* Natural genetic variation in the pheromone production of *C. elegans*. *Proc Natl Acad Sci U S A.* 2023:120:e2221150120. https://doi.org/10.1073/pnas.2221150120.

Li H. 2013 May 26. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM [preprint]. arXiv [q-bio.GN]. http://arxiv.org/abs/1303.3997

Li WH. Unbiased estimation of the rates of synonymous and nonsynonymous substitution. *J Mol Evol*. 1993:36:96–99. https://doi.org/10.1007/BF02407308.

Li L, Comi TJ, Bierman RF, Akey JM. Recurrent gene flow between Neanderthals and modern humans over the past 200,000 years. *Science*. 2024:385:eadi1768. https://doi.org/10.1126/science.adi1768.

Llorens C *et al.* The Gypsy Database (GyDB) of mobile genetic elements: release 2.0. *Nucleic Acids Res*. 2011:39:D70–D74. https://doi.org/10.1093/nar/gkq1061.

Machado HE *et al.* Comparative population genomics of latitudinal variation in *Drosophila simulans* and *Drosophila melanogaster*. *Mol Ecol*. 2016:25:723–740. https://doi.org/10.1111/mec.13446.

Marçais G *et al.* MUMmer4: a fast and versatile genome alignment system. *PLoS Comput Biol*. 2018:14:e1005944. https://doi.org/10.1371/journal.pcbi.1005944.

Minh BQ *et al.* IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Mol Biol Evol.* 2020:37:1530–1534. https://doi.org/10.1093/molbev/msaa015.

Mistry J *et al.* Pfam: the protein families database in 2021. *Nucleic Acids Res*. 2021:49:D412–D419. https://doi.org/10.1093/nar/gkaa913.

Moya ND *et al.* Novel and improved *Caenorhabditis briggsae* gene models generated by community curation. *BMC Genomics*. 2023:24:486. https://doi.org/10.1186/s12864-023-09582-0.

Moya ND, Yan SM, McCoy RC, Andersen EC. The long and short of hyperdivergent regions. *Trends Genet*. 2025:41:303–314. https://doi.org/10.1016/j.tig.2024.11.005.

Noble LM *et al.* Selfing is the safest sex for *Caenorhabditis tropicalis*. *elife*. 2021:10:e62587. https://doi.org/10.7554/eLife.62587.

Nordborg M, Donnelly P. The coalescent process with selfing. *Genetics*. 1997:146:1185–1195. https://doi.org/10.1093/genetics/146.3.1185.

Ortiz EM. 2019. *vcf2phylip v2.0: convert a VCF matrix into several matrix formats for phylogenetic analysis*. Zenodo. https://zenodo.org/record/1257057

Patterson N, Price AL, Reich D. Population structure and eigenanalysis. *PLoS Genet*. 2006:2:e190. https://doi.org/10.1371/journal.pgen.0020190.

Pedersen BS, Quinlan AR. Mosdepth: quick coverage calculation for genomes and exomes. *Bioinformatics*. 2018:34:867–868. https://doi.org/10.1093/bioinformatics/btx699.

Pertea G, Pertea M. GFF utilities: GffRead and GffCompare. *F1000Res*. 2020:9:ISCB Comm J-304. https://doi.org/10.12688/f1000research.23297.2.

Poplin R, *et al.* 2018 July 24. Scaling accurate genetic variant discovery to tens of thousands of samples [preprint]. bioRxiv 201178. https://www.biorxiv.org/content/10.1101/201178v3

Price AL *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet*. 2006:38:904–909. https://doi.org/10.1038/ng1847.

Purcell S *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet*. 2007:81:559–575. https://doi.org/10.1086/519795.

Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*. 2010:26:841–842. https://doi.org/10.1093/bioinformatics/btq033.

Reilly PF, Tjahjadi A, Miller SL, Akey JM, Tucci S. The contribution of Neanderthal introgression to modern human traits. *Curr Biol*. 2022:32:R970–R983. https://doi.org/10.1016/j.cub.2022.08.027.

Rockman MV, Kruglyak L. Recombinational landscape and population genomics of *Caenorhabditis elegans*. *PLoS Genet*. 2009:5:e1000419. https://doi.org/10.1371/journal.pgen.1000419.

Rockman MV, Tintori SC, Nguyen THM, Yomai VMH. Caenorhabditis diversity on Pohnpei, Micronesia, provides evidence that the Elegans Supergroup has its roots in the Americas and diversified in the Pacific en route to Asia. *Evolution*. 2026:80:1013–1034. https://doi.org/10.1093/evolut/qpag031.

Rockman MV, Skrovanek SS, Kruglyak L. Selection at linked sites shapes heritable phenotypic variation in *C. elegans*. *Science*. 2010:330:372–376. https://doi.org/10.1126/science.1194208.

Rognes T, Flouri T, Nichols B, Quince C, Mahé F. VSEARCH: a versatile open source tool for metagenomics. *PeerJ*. 2016:4:e2584. https://doi.org/10.7717/peerj.2584.

Saxena AS, Salomon MP, Matsuba C, Yeh S-D, Baer CF. Evolution of the mutational process under relaxed selection in *Caenorhabditis elegans*. *Mol Biol Evol*. 2019:36:239–251. https://doi.org/10.1093/molbev/msy213.

Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*. 2015:31:3210–3212. https://doi.org/10.1093/bioinformatics/btv351.

Smit AFA, Hubley R, Green P. RepeatMasker Open-4.0. RepeatMasker Open-4.0; 2015. http://www.repeatmasker.org

Städler T, Haubold B, Merino C, Stephan W, Pfaffelhuber P. The impact of sampling schemes on the site frequency spectrum in nonequilibrium subdivided populations. *Genetics*. 2009:182:205–216. https://doi.org/10.1534/genetics.108.094904.

Stevens L *et al.* Chromosome-level reference genomes for two strains of *Caenorhabditis briggsae*: an improved platform for comparative genomics. *Genome Biol Evol*. 2022:14:evac042. https://doi.org/10.1093/gbe/evac042.

Stevens L *et al.* Ancient diversity in host-parasite interaction genes in a model parasitic nematode. *Nat Commun*. 2023:14:7776. https://doi.org/10.1038/s41467-023-43556-w.

Stockdale C, Swiderski MR, Barry JD, McCulloch R. Antigenic variation in *Trypanosoma brucei*: joining the DOTs. *PLoS Biol*. 2008:6:e185. https://doi.org/10.1371/journal.pbio.0060185.

Suyama M, Torrents D, Bork P. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res*. 2006:34:W609–W612. https://doi.org/10.1093/nar/gkl315.

Tajima F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*. 1989:123:585–595. https://doi.org/10.1093/genetics/123.3.585.

Takahata N, Nei M. Allelic genealogy under overdominant and frequency-dependent selection and polymorphism of major histocompatibility complex loci. *Genetics*. 1990:124:967–978. https://doi.org/10.1093/genetics/124.4.967.

Tarasov A, Vilella AJ, Cuppen E, Nijman IJ, Prins P. Sambamba: fast processing of NGS alignment formats. *Bioinformatics*. 2015:31:2032–2034. https://doi.org/10.1093/bioinformatics/btv098.

Teterina AA *et al.* Genomic diversity landscapes in outcrossing and selfing Caenorhabditis nematodes. *PLoS Genet*. 2023:19:e1010879. https://doi.org/10.1371/journal.pgen.1010879.

Thomas CG *et al.* Full-genome evolutionary histories of selfing, splitting, and selection in Caenorhabditis. *Genome Res*. 2015:25:667–678. https://doi.org/10.1101/gr.187237.114.

UniProt Consortium. UniProt: the universal protein knowledgebase in 2025. *Nucleic Acids Res*. 2025:53:D609–D617. https://doi.org/10.1093/nar/gkae1010.

Walker BJ *et al.* Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One*. 2014:9:e112963. https://doi.org/10.1371/journal.pone.0112963.

Watterson GA. On the number of segregating sites in genetical models without recombination. *Theor Popul Biol.* 1975:7:256–276. https://doi.org/10.1016/0040-5809(75)90020-9.

Wong TKF *et al.* MAST: phylogenetic inference with mixtures across sites and trees. *Syst Biol*. 2024:73:375–391. https://doi.org/10.1093/sysbio/syae008.

Wong TKF *et al.* IQ-TREE 3: phylogenomic inference software using complex evolutionary models. *Mol Biol Evol*. 2026:43:msag117. https://doi.org/10.1093/molbev/msag117.

Woodruff GC, Eke O, Baird SE, Félix M-A, Haag ES. Insights into species divergence and the evolution of hermaphroditism

from fertile interspecies hybrids of Caenorhabditis nematodes. *Genetics*. 2010:186:997–1012. https://doi.org/10.1534/genetics.110.120550.

Wu T *et al.* clusterProfiler 4.0: a universal enrichment tool for interpreting omics data. *Innovation (Camb.)*. 2021:2:100141. https://doi.org/10.1016/j.xinn.2021.100141.

Xu S *et al.* Using clusterProfiler to characterize multiomics data. *Nat Protoc*. 2024:19:3292–3320. https://doi.org/10.1038/s41596-024-01020-z.

Yang X *et al.* Comparative population genomics reveals convergent and divergent selection in the apricot-peach-plum-mei complex. *Hortic Res.* 2024:11:uhae109. https://doi.org/10.1093/hr/uhae109.

Yan SM *et al.* Local adaptation and archaic introgression shape global diversity at human structural variant loci. *eLife*. 2021:10:e67615. https://doi.org/10.7554/eLife.67615.

Yu G. Thirteen years of clusterProfiler. *Innovation (Camb.)*. 2024:5:100722. https://doi.org/10.1016/j.xinn.2024.100722.

Yu G, Wang L-G, Han Y, He Q-Y. clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS*. 2012:16:284–287. https://doi.org/10.1089/omi.2011.0118.

Yu G, Smith DK, Zhu H, Guan Y, Lam TT-Y. Ggtree: an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods Ecol Evol*. 2017:8:28–36. https://doi.org/10.1111/2041-210X.12628.