Chromosome-Level Reference Genomes for Two Strains of *Caenorhabditis briggsae*: An Improved Platform for Comparative Genomics

Lewis Stevens [ID] [1,1], Nicolas D. Moya[1,2], Robyn E. Tanny[1], Sophia B. Gibson [ID] [1], Alan Tracey[3], Huimin Na[4], Rojin Chitrakar[5], Job Dekker[4], Albertha J.M. Walhout[4], L. Ryan Baugh[5,6], and Erik C. Andersen [ID] [1,*]

[1]Department of Molecular Biosciences, Northwestern University, Evanston, Illinois, USA

[2]Interdisciplinary Biological Sciences Program, Northwestern University, Evanston, Illinois, USA

[3]Tree of Life, Wellcome Sanger Institute, Cambridge, UK

[4]Department of Systems Biology, University of Massachusetts Chan Medical School, Worcester, Massachusetts, USA

[5]Department of Biology, Duke University, Durham, North Carolina, USA

[6]Center for Genomic and Computational Biology, Duke University, Durham, North Carolina, USA

*Corresponding author: E-mail: erik.andersen@northwestern.edu.

[1]Present address: Tree of Life, Wellcome Sanger Institute, Cambridge, UK.

Abstract

The publication of the *Caenorhabditis briggsae* reference genome in 2003 enabled the first comparative genomics studies between *C. elegans* and *C. briggsae*, shedding light on the evolution of genome content and structure in the *Caenorhabditis* genus. However, despite being widely used, the currently available *C. briggsae* reference genome is substantially less complete and structurally accurate than the *C. elegans* reference genome. Here, we used high-coverage Oxford Nanopore long-read and chromosome-conformation capture data to generate chromosome-level reference genomes for two *C. briggsae* strains: QX1410, a new reference strain closely related to the laboratory AF16 strain, and VX34, a highly divergent strain isolated in China. We also sequenced 99 recombinant inbred lines generated from reciprocal crosses between QX1410 and VX34 to create a recombination map and identify chromosomal domains. Additionally, we used both short- and long-read RNA sequencing data to generate high-quality gene annotations. By comparing these new reference genomes to the current reference, we reveal that hyper-divergent haplotypes cover large portions of the *C. briggsae* genome, similar to recent reports in *C. elegans* and *C. tropicalis*. We also show that the genomes of selfing *Caenorhabditis* species have undergone more rearrangement than their outcrossing relatives, which has biased previous estimates of rearrangement rate in *Caenorhabditis*. These new genomes provide a substantially improved platform for comparative genomics in *Caenorhabditis* and narrow the gap between the quality of genomic resources available for *C. elegans* and *C. briggsae*.

Key words: *Caenorhabditis briggsae*, reference genomes, comparative genomics, selfing, genetic diversity, genome rearrangement.

Significance

Despite the long history of using the free-living nematode *Caenorhabditis briggsae* alongside the model organism *C. elegans* to understand the evolution of *Caenorhabditis* development and genetics, the genomic resources available for *C. briggsae* lag behind those available for *C. elegans*. In this study, we used the latest sequencing technologies to generate high-quality reference genomes for two strains of *C. briggsae* and used these new genomes to study within-species divergence and compare rates of genome rearrangement between selfing and outcrossing *Caenorhabditis* species. As demonstrated by the analyses we present, our new resources provide a substantially improved platform for comparative genomics between *C. briggsae* and *C. elegans*.

Introduction

Since its introduction as a model for animal development in the 1950s, the free-living nematode *Caenorhabditis elegans* has become one of the key model organisms in biology. However, before eventually settling on *C. elegans*, Sydney Brenner had initially proposed a closely related species, *Caenorhabditis briggsae*, for study in the laboratory (Edgar and Wood 1977; Riddle et al. 2011). Although *C. briggsae* has received far less attention than its close relative, it is now widely used as a satellite model organism and has played a key role in our understanding of *Caenorhabditis* evolution, development, and genetics (Baird and Chamberlin 2006). *Caenorhabditis elegans* and *C. briggsae* share many features of their biology: they are nearly morphologically indistinguishable; both species reproduce predominantly by self-fertilization; and both are found globally, often in the same local habitats (Nigon and Dougherty 1949; Barrièreand Félix 2005; Cutter et al. 2006; Dolgin et al. 2007; Félix and Duveau 2012; Thomas et al. 2015; Crombie et al. 2019). However, it is their differences that have helped shed light on several key biological pathways. For example, *C. elegans* and *C. briggsae* independently evolved from outcrossing species to reproduce as selfing hermaphrodites (which are essentially females capable of producing sperm) and have done so using distinct genetic mechanisms (Cho et al. 2004; Nayak et al. 2005; Hill et al. 2006; Kiontke et al. 2011).

In 2003, five years after the *C. elegans* reference genome was published (*C. elegans* Sequencing Consortium 1998), a reference genome for the laboratory strain of *C. briggsae*, AF16, was generated using Sanger-based shotgun sequencing and a physical map generated using fosmids and bacterial artificial chromosomes (BACs) (Stein et al. 2003). This genome was subsequently resolved into chromosomes using a genetic map constructed by sequencing lines from interstrain crosses (Hillier et al. 2007; Ross et al. 2011). The availability of a high-quality reference genome for *C. briggsae* enabled the first comparative genomics studies between *C. elegans* and *C. briggsae*, revealing that their genomes have diverged substantially since they last shared a common ancestor, and have undergone a strikingly high rate of intrachromosomal rearrangement (Coghlan and Wolfe 2002). Despite these differences, the higher-order structure of the *C. elegans* genome is largely conserved in *C. briggsae* and surprisingly few genes have moved between chromosomes (Hillier et al. 2007). The reference genome also provided a foundation for population genetic and genomic studies, revealing that *C. briggsae* harbors higher levels of genetic diversity than *C. elegans*, largely driven by the existence of distinct phylogeographic groups that makes *C. briggsae* a more suitable model for studying gene flow and speciation (Thomas et al. 2015). Nearly 20 years later, the *C. briggsae* reference genome remains one of the highest quality *Caenorhabditis* genomes currently available (Harris et al. 2020). However, it still lags behind the *C. elegans* genome in terms of accuracy and completeness, containing thousands of gaps that are estimated to comprise several megabases of sequence along with numerous regions that have been found to be misoriented (Ren et al. 2018).

Since the *C. briggsae* reference genome was published, DNA sequencing technology has advanced at an exponential rate. In recent years, long-read sequencing technologies, namely those offered by Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT), have revolutionized genome assembly, and it is now relatively straightforward to generate highly contiguous genome assemblies (Rhie et al. 2021). Moreover, high-throughput mapping technologies, such as chromosome-conformation capture (Hi-C), are now routinely being used to construct fully chromosome-level reference genomes that far surpass the quality of their predecessors (Rhie et al. 2021). Similarly, it is now possible to generate long-read RNA sequencing (RNA-seq) data to assemble full-length transcripts and substantially improve the quality of automated gene annotations. Substantially improved reference genomes generated using these technologies have recently been published for several nematode species, including for *Caenorhabditis tropicalis* and *Caenorhabditis remanei* (Teterina et al. 2020; Noble et al. 2021), *Oscheius tipulae* (Gonzalez de la Rosa et al. 2021), and the ruminant parasite *Haemonchus contortus* (Doyle et al. 2020). These

Table 1

Reference genome metrics

	C. briggsae QX1410	*C. briggsae* VX34	*C. briggsae* AF16	*C. elegans* N2
Accession	PRJNA784955	PRJNA784955	PRJNA10731	PRJNA13758
Version	v1	v1	WS279	WS279
Span (Mb)	106.2	107.0	108.4	100.2
Number of scaffolds	6 (+MT)	6 (+MT)	367	6 (+MT)
Number of unassigned scaffolds	0	0	361	0
Percentage of assembly span in in six scaffolds (%)	100	100	97.0	100
Scaffold N50 (Mb)	17.1	17.3	17.5	17.5
Number of contigs[a]	14	10	5,074	6
Contig N50 (Mb)[a]	14.7	16.1	0.05	17.5
Number of gaps	7	3	4,707	0
Span of Ns (kb)	3.5	1.5	2,965.5	0
BUSCO[b] completeness (%)	99.4	99.4	98.8	99.4
QV score[c]	45.6	44.4	–	–

[a]Contig values calculated by splitting scaffolds at ≥10 consecutive Ns.
[b]Genome completeness was assessed using BUSCO (version 4.1.4) with the nematoda_odb10 dataset.
[c]QV scores were calculated by Merqury (version 1.1) using short-read Illumina data. QV scores of 45.6 and 44.4 correspond to one error every 36.2 and 27.5 kb, respectively.

technologies have even been used to identify and extend multiple repetitive regions that are collapsed in the *C. elegans* N2 reference genome (Tyson et al. 2018; Yoshimura et al. 2019), widely regarded as one of the highest quality eukaryotic reference genomes available.

Here, we use the Oxford Nanopore PromethION platform and Hi-C to sequence the genomes of two *C. briggsae* strains, QX1410, a new reference strain closely related to AF16, and VX34, a highly divergent strain. QX1410 was chosen instead of improving the AF16 genome because the fidelity of the AF16 strain is no longer certain after its long time in the laboratory. We use these data to generate two high-quality reference genomes that have biological completeness scores equal to that of the current *C. elegans* reference genome, and they have substantially fewer gaps and unplaced sequence than the existing *C. briggsae* AF16 reference genome. We also use both long- and short-read RNA-seq data to generate high-quality gene annotations. We also generated and sequenced a panel of recombinant inbred lines (RILs) from reciprocal crosses between QX1410 and VX34 to characterize the recombination landscape of the *C. briggsae* genome. Consistent with previous reports in *C. elegans* and *C. tropicalis*, we find that hyper-divergent haplotypes cover large portions of the *C. briggsae* genome. We also revisit one of the first comparative genomics analyses performed between *C. briggsae* and *C. elegans* and reveal that the genomes of selfing *Caenorhabditis* species have undergone more rearrangement than their outcrossing relatives, which has biased previous estimates of rearrangement rate in *Caenorhabditis*. These new resources will provide a substantially improved foundation for genomic analyses in this important satellite model organism.

Results

High-Quality Reference Genomes for Two *C. briggsae* Strains

We sought to generate a high-quality reference genome for QX1410, a new reference strain for *C. briggsae* isolated from Saint Lucia that is closely related to AF16, the current reference strain (Thomas et al. 2015). To facilitate comparative analyses, we also sequenced the genome of a highly divergent strain, VX34, isolated from China. We generated high-coverage long-read data for each strain (read length N50s of 23.6 and 23.5 kb, respectively; coverages of 219× and 508×, respectively) using the Oxford Nanopore PromethION platform and sequenced chromosome-conformation capture (Hi-C) libraries for each strain to high coverage (173× and 168×, respectively) using Illumina technology. We assembled the long PromethION reads independently for each strain using several tools and chose the most contiguous assemblies, both of which comprised several contigs that represented complete chromosomes (supplementary fig. S1, Supplementary Material online). We corrected sequencing errors in the contigs, and these polished contigs were then scaffolded into complete chromosomes using the Hi-C data (supplementary fig. S2, Supplementary Material online). Each reference genome was manually curated by inspecting coverage of the long-reads and by assessing congruence with the Hi-C contact maps.

The resulting reference genomes for QX1410 and VX34 span 106.2 and 107.0 Mb, respectively, similar in span to the existing AF16 reference (table 1). Both comprise six scaffolds representing the six chromosomes (I–V, X), a complete mitochondrial genome, and have no unplaced

sequence. The QX1410 and VX34 assemblies have high base-pair accuracy, with consensus quality value (QV) scores of 45.6 (which corresponds to one error in 36.2 kb) and 44.4 (one error in 27.5 kb), respectively. BUSCO completeness is equal to that of the current *C. elegans* N2 reference genome (99.4%), compared with 98.8% for the existing *C. briggsae* AF16 reference (table 1). The new reference genomes are substantially more contiguous than AF16, with contig N50s of 14.7 and 16.1 Mb, respectively, compared with 47 kb. Only seven gaps remain in the QX1410 reference genome (one in chromosome I, one in chromosome II, and five in the X chromosome) and three gaps in the VX34 reference genome (two in chromosome in V and one in the X chromosome) compared with 4,706 gaps in the existing AF16 reference genome. In contrast to the chromosomal scaffolds in AF16, all of which lack the nematode telomeric repeat sequence ([TTAGGC]*n*) at their ends, a majority of the chromosomes in our reference genomes ends in telomeric repeat sequences (supplementary fig. S3, Supplementary Material online).

Despite having high-coverage long-read data, manual curation revealed that the subtelomeric regions, which are known to be highly repetitive in *C. elegans* (Kim et al. 2019), are unresolved in five of the 12 ends of the QX1410 reference genome. One of these is the left-end of chromosome V (VL), which ends in nine tandemly repeated ~7.5 kb ribosomal DNA (rDNA) cistron units (supplementary fig. S4A, Supplementary Material online). In *C. elegans*, the rDNA cluster consists of ~55 repeated units and sits adjacent to the telomeric repeat sequence on chromosome I (Ellis et al. 1986; *C. elegans* Sequencing Consortium 1998). In QX1410, the average coverage within this repeated region is 943×, ~6.5-times higher than the chromosome-wide average (~146×), suggesting that this region is collapsed in our assembly and that the true number of rDNA repeats in QX1410 is ~58 (supplementary fig. S4B, Supplementary Material online). Another unresolved region is the right end of the X chromosome (XR), which ends in a large ~65 kb tandem repeat (supplementary fig. S5A, Supplementary Material online) that presumably prevented the genome assembler from extending into the telomeric repeat sequence. Two of the remaining unresolved chromosome ends, IL and IVL, end in unique sequences that are punctuated with blocks of kmers that contain the nematode telomeric sequence. For example, in the IVL subtelomere, we find multiple blocks of a repeating 19-mer (TTAGGCTTAGGCTTCCCGC) interspersed with a unique sequence (supplementary figs. S5B and S6A, Supplementary Material online). Blocks of the same sequence are also found in the IVR subtelomere (supplementary fig. S6A, Supplementary Material online). In the IL subtelomere, we find a block of a repeating 14-mer (TAAGCCTAAGCCTC) with blocks of the same

sequence also found on the IIR and XL (supplementary fig. S6A, Supplementary Material online). Similar blocks of both kmers exist in the subtelomeric regions of the VX34 genome, but they are found at subtelomeres of different chromosomes (supplementary fig. S6B, Supplementary Material online).

The AF16 reference genome has previously been reported to contain multiple mis-scaffolded regions (Ren et al. 2018). Although efforts have been made to correct these regions, these changes are not yet reflected in the current version of the reference available via public databases (Harris et al. 2020). To assess collinearity between the three *C. briggsae* reference genomes, we aligned AF16 and VX34 to the QX1410 reference genome. Consistent with previous reports, we found numerous regions in all six chromosomes that are inverted in AF16 relative to QX1410 (fig. 1C). By contrast, and despite being far more divergent from QX1410 than AF16 and being assembled independently, the VX34 genome is highly collinear with the QX1410 reference genome (fig. 1B). To quantify this relationship, we called structural variants in the reference genomes of AF16 and VX34 using QX1410 as the reference and focused on large inversions (≥50 kb in length). Relative to QX1410, AF16 has 47 inversions (average size of 167 kb) and VX34 has six (average 83 kb). Of those inversions above 100 kb, AF16 has 31 and VX34 has just one. Importantly, the QX1410 and VX34 reference genomes were generated by scaffolding contigs that spanned multiple megabases (contig N50s of 14.7 and 16.1 Mb, respectively), meaning that all of the inversions called in VX34 fall within contiguous sequence and are thus likely to represent real structural variation. In contrast, the AF16 reference genome was scaffolded using a highly fragmented assembly (contig N50 of 47 kb), meaning many of the observed variants are likely to be scaffolding errors. In summary, the new reference genomes that we have generated for QX1410 and VX34 are substantially more contiguous, complete, and structurally correct than the current AF16 reference genome.

High-Quality Protein-Coding Gene Annotations Using Long- and Short-Read RNA-Seq

We sought to develop a computational pipeline that leverages long- and short-read RNA-seq reads to generate high-quality protein-coding gene annotations for the QX1410 and VX34 reference genomes. We collected RNA from mixed-stage, male-enriched, and starved cultures to maximize transcript detection. To allow us to use the highly curated *C. elegans* N2 reference annotation (PRJNA13758) as a truth set, we sequenced the transcriptome of *C. elegans* strain PD1074 (a recent clone of N2) using the Pacific Biosciences (PacBio) Single-Molecule Real-Time (SMRT) platform and benchmarked several transcriptome assemblies
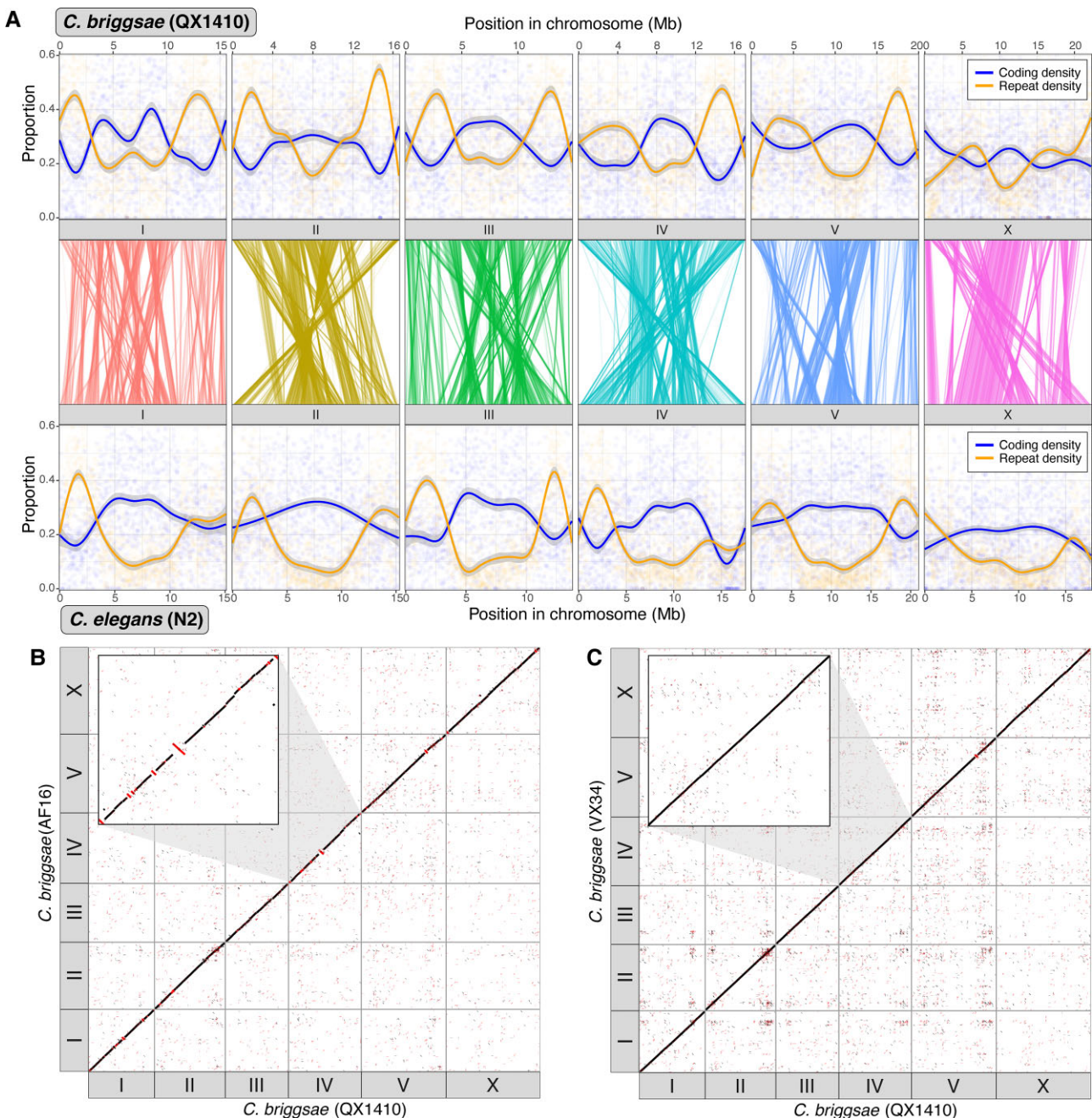
FIG. 1.—High-quality reference genomes for two *C. briggsae* strains. (*A*) Comparison between the *C. briggsae* QX1410 and *C. elegans* N2 reference genomes. Repeat density and protein-coding gene density per 10 kb windows are shown. Repeats were identified *de novo* using RepeatModeler2. Solid lines represent LOESS smoothing functions fitted to the data. Relative positions of 10,387 one-to-one orthologs are shown as lines joining the two density plots. (*B*) Whole-genome alignment of AF16 to the QX1410 reference genome generated using nucmer. Alignments shorter than 1 kbp are not shown. Alignments in the reverse orientation are highlighted in red. Inset: chromosome IV showing multiple regions between AF16 and QX1410 that are in different orientations. (*C*) Whole-genome alignment of VX34 to the QX1410 reference genome generated using nucmer. Alignments shorter than 1 kbp are not shown. Alignments in the reverse orientation are highlighted in red. Inset: the same chromosome IV region as in (*B*) showing a largely collinear alignment.

and gene prediction tools. We refined the PacBio long-reads into 55,936 high-quality transcripts and generated gene models by predicting open reading frames (ORFs) in the high-quality transcripts. Additionally, we generated protein-coding gene predictions from short RNA-seq read

alignments. Based on our N2 benchmark, we merged the best transcriptome assembly and gene prediction models into a single, nonredundant gene set. The BUSCO completeness of our gene set was 99.4%, only slightly lower than the completeness of the existing *C. elegans* N2 reference gene

set (100%). Merging the short- and long-read-based gene models led to improvements in exon, intron, and transcript sensitivity relative to either approach alone (supplementary table S1, Supplementary Material online). Our final annotation has correct predictions for 80.7% and 90.7% of all *C. elegans* exons and introns, respectively. We correctly predicted 53.1% of all *C. elegans* transcripts, with at least one correctly predicted transcript in 75.8% of the *C. elegans* genes.

Using this same process, we sequenced the transcriptomes of both QX1410 and VX34 strains using PacBio SMRT and Illumina platforms and employed our pipeline to generate high-quality gene models for both *C. briggsae* strains. We predicted 19,947 and 19,737 protein-coding genes in the QX1410 and VX34 genomes, respectively, similar to the number of genes predicted for AF16 (20,821). BUSCO completeness scores of the QX1410 and VX34 annotations are 99.4% and 99.3%, respectively, marginally higher than the current AF16 (WS280) annotation (99.2%), which has undergone extensive manual curation. We also assessed gene annotation quality by comparing the protein sequence length of each gene in our gene sets to their corresponding orthologs in *C. elegans* (supplementary fig. S7, Supplementary Material online). We performed the same protein-length analysis for the current *C. briggsae* AF16 reference annotation (WS280) and a version of the AF16 annotation composed of automated predictions only (WS255). Our QX1410 and VX34 annotations show substantial improvements in protein-length accuracy relative to the uncurated AF16 WS255 release (supplementary table S2, Supplementary Material online). However, compared with the AF16 WS280 release, we find that our protein sequences have fewer matches that are identical in length or within 5% of the length of the orthologous *C. elegans* protein (supplementary table S2, Supplementary Material online). This result suggests that the manual curation of the AF16 gene set, which included the use of *C. elegans* protein alignments, has corrected structural errors that were present in the automated gene set (WS255). Manual curation is therefore needed to further improve the QX1410 and VX34 gene models to bring these annotations to protein-length accuracies comparable to the manually curated AF16 WS280 release.

Recombination in the *C. briggsae* Recombinant Lines

To characterize the recombination landscape between QX1410 and VX34, we generated a panel of 99 F2 RILs from reciprocal crosses between the two strains and genotyped 2,981 single nucleotide markers in all RILs. After removing markers with distorted segregation patterns and markers with large deviations in frequency relative to neighbors (supplementary fig. S8 and table S3, Supplementary Material online), the remaining 2,828 markers were used to

estimate a genetic map for *C. briggsae*. The estimated genetic map has a length of 508.8 cM and spans 97.7% of the QX1410 genome physical length. The size of this genetic map is similar to that of previous estimations using an F2 RIL scheme (588.1 cM) (Hillier et al. 2007) and ~55% of the size of previous estimations using an advanced intercross (AI) scheme (928.6 cM) (Ross et al. 2011). The reduction in size relative to this previous map can be explained, at least in part, by the reduced number of recombination breakpoints expected in RILs relative to recombinant lines from AI (Rockman and Kruglyak 2008, 2009).

As in *C. elegans*, the genetic maps for the six *C. briggsae* chromosomes have distinct arms and centers that show detectable changes in recombination rate (Hillier et al. 2007; Rockman and Kruglyak 2009; Ross et al. 2011). We generated Marey maps to show the genetic position as a function of the physical position and used segmented linear regression to identify arm-center boundaries and estimate the rate of recombination in each domain (fig. 2A, table 2). Most of our domain boundaries are in agreement with the span and physical position of previously defined domains (Ross et al. 2011). Although the recombination rate is relatively constant within each domain, we observed small segments in chromosomal arms where recombination rate abruptly approached zero. This pattern can be explained by sampling error, where the limited number of recombination events that occur across the 99 RILs could lead to several genomic markers with skewed recombination fractions. With the exception of IIL, the recombination rate at chromosome ends sharply decreased, as expected for regions approaching chromosome arm-tip boundaries. These RILs clearly resolve tip domains at IIIR, IVL, IVR, and XL. Tip domain boundaries were manually defined.

In selfing *Caenorhabditis* species, evidence suggests that selective pressure to maintain the linkage between coadapted alleles might lead to incompatibilities between interbreeding populations (Seidel et al. 2008; Rockman and Kruglyak 2009; Ross et al. 2011; Noble et al. 2021). A strong departure from the neutral expectation in allele frequencies was observed in chromosomes I, II, and IV (fig. 2B). These skews might be explained by regions of incompatibility between QX1410 and VX34. Interestingly, the skew in chromosome I spans the entirety of the chromosome center. Although this skew may be the result of massive chromosome I incompatibilities between the two strains, markers in chromosome I center are sparse. Incompatibilities surrounding these sparse markers might drive inflation in average allele frequencies across this genomic interval.

High Divergence Among the *C. briggsae* Reference Genomes

The genomes of *C. elegans* wild isolates contain large, hyper-divergent haplotypes that comprise unique sets of
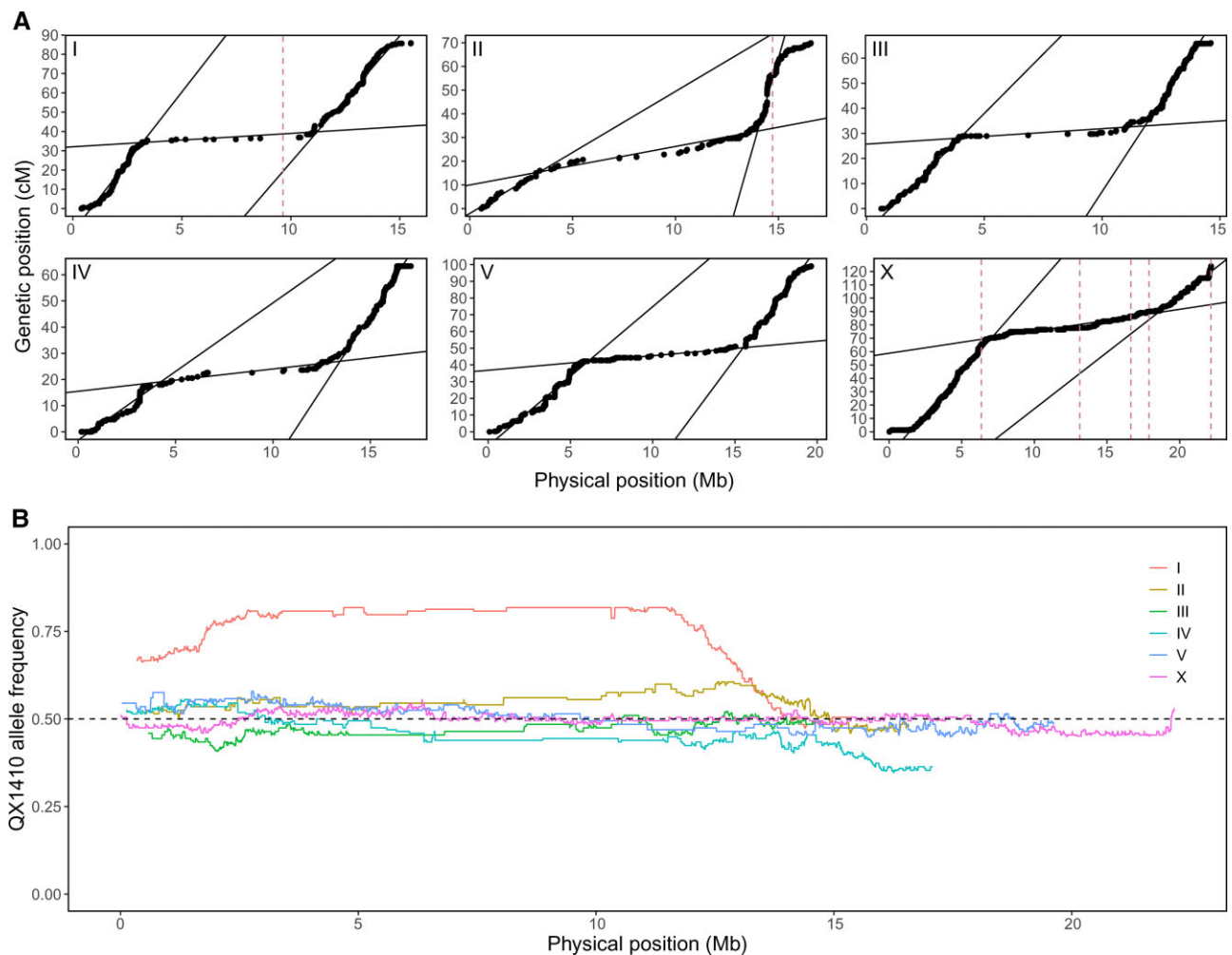
Fig. 2.—Recombination rates in the *C. briggsae* genome determined by genotyping 99 QX1410xVX34 RILs. (*A*) Marey maps for each chromosome in QX1410. The genetic position of each marker is shown as a function of physical position (black dots). Fits from segmented linear regressions are shown as black lines. Changepoints in the segmented linear regressions were used to estimate chromosome domain boundaries and the rate of recombination. Dashed pink lines indicate the physical position of gaps in the QX1410 genome assembly. Asterisks were added to chromosomal ends where subtelomeric regions are unresolved. (*B*) Frequency of the QX1410 allele as a function of physical position across every marker in each chromosome. Allele frequency was averaged using a sliding window 100 kb with a step size of 5 kb. The neutral expected frequency of 0.5 is shown as a dashed horizontal black line.

genes and alleles that are highly diverged at the amino acid level (Thompson et al. 2015; Lee et al. 2021). These haplotypes are hypothesized to be remnants of genetic diversity present in the outcrossing ancestor that has been maintained by long-term balancing selection since the evolution of selfing (Thompson et al. 2015; Lee et al. 2021). Similar hyper-divergent regions have been reported in both *C. briggsae* (Lee et al. 2021) and *C. tropicalis* another related selfing *Caenorhabditis* species (Ben-David et al. 2021; Noble et al. 2021). To quantify the genome-wide divergence between the three sequenced *C. briggsae* strains, we aligned the AF16 and VX34 reference genomes to QX1410, calculated alignment identity, and called variants (SNVs and indels). Given that hyper-divergent haplotypes in *C. elegans* often show little homology to each other at the

nucleotide level, we also identified alleles across all three strains using an orthology clustering approach and calculated amino acid identity.

Consistent with recent findings in all other selfing *Caenorhabditis* species (Lee et al. 2021; Noble et al. 2021), we find that hyper-divergent haplotypes are widespread across the genomes of all three *C. briggsae* strains (fig. 3). Despite QX1410 and AF16 belonging to the same "Tropical" genetic group (Thomas et al. 2015), large regions of their genomes are unalignable at the nucleotide level (fig. 3*A*). As in *C. elegans*, these regions are overrepresented in the autosomal arms and underrepresented in the autosomal centers and on the X chromosome (Lee et al. 2021). The nucleotide divergence in these regions is associated with a substantial drop in amino acid identity

Table 2

Chromosomal domains

Chr		Left tip	Left arm	Center	Right arm	Right tip
I	Size (kb)	388	2,803	8,017	3,762	571
	Size (%)	2.5	18.0	51.6	24.2	3.7
	Right end (kb)	388	3,191	11,208	14,970	15,541
	Rate[a] (cM/Mb)	0	8.43	0.40	7.63	0
II	Size (kb)	566	2,844	10,578	2,583	24
	Size (%)	3.4	17.1	63.7	15.6	0.2
	Right end (kb)	566	3,410	13,988	16,571	16,595
	Rate[a] (cM/Mb)	0	3.68	1.17	21.75	0
III	Size (kb)	716	3,249	7,859	2,171	816
	Size (%)	4.8	21.9	53.1	14.7	5.5
	Right end (kb)	716	3,965	11,824	13,995	14,811
	Rate[a] (cM/Mb)	0	7.16	0.47	10.88	0
IV	Size (kb)	702	3,546	9,194	2,924	714
	Size (%)	4.1	20.8	53.8	17.1	4.2
	Right end (kb)	702	4,248	13,442	16,366	17,080
	Rate[a] (cM/Mb)	0	4.14	0.67	8.99	0
V	Size (kb)	469	5,593	9,433	4,052	386
	Size (%)	2.4	28.1	47.3	20.3	1.9
	Right end (kb)	469	6,062	15,495	19,547	19,933
	Rate[a] (cM/Mb)	0	4.20	0.44	6.68	0
X	Size (kb)	1,444	5,622	11,456	3,639	60
	Size (%)	6.5	25.3	51.5	16.4	0.3
	Right end (kb)	1,444	7,066	18,522	22,161	22,221
	Rate[a] (cM/Mb)	0	5.01	0.67	3.46	0
Cumulative size (kb)		4,285	24,385	55,809	19,131	2,571
Cumulative size (%)		4.0	23.0	52.6	18.0	2.4

[a]Rates are estimated from the slopes of segmented linear fits with chromosome genetic length scaled to 50 cM.

between alleles, with many alleles showing <95% identity (fig. 3A). For example, large regions of the right arm of chromosome II are unalignable between QX1410 and AF16, and alleles across this arm have a mean amino acid identity of 98.6% (fig. 3A). Comparing QX1410 with the more divergent VX34, we find a substantial divergence in the arms of all chromosomes (fig. 3B). As in AF16, the right arms of chromosome II and V are particularly divergent, with a mean amino acid identity of 98.4% and 98.0% between alleles, respectively (fig. 3B). Surprisingly, we also find substantial divergence on the X chromosome between QX1410 and VX34, with a mean amino acid identity of 98.9% across its length and large sections of the chromosome arms show poor or no alignment (fig. 3B).

To place the level of divergence in *C. briggsae* in the context of previous work in *C. elegans*, we also compared aligned long-read assemblies for CB4856 (a divergent strain from Hawaii) and XZ1516 (the most divergent *C. elegans* strain currently known, also isolated in Hawaii) to the laboratory N2 reference genome and called variants (SNVs and indels). As expected, the divergence between the QX1410 and AF16 is the lowest of all the comparisons, with one SNV every 634 bp. The divergence between QX1410 and VX34 (average of one SNV every 135 bp) is substantially higher than the higher divergence between the *C. elegans* strains N2 and CB4856 (one SNV every 362 bp, respectively) (supplementary fig. S9A–C, Supplementary Material online) and similar to that between N2 and the most divergent *C. elegans* strain XZ1516 (one SNV in 141 bp) (supplementary fig. S9D, Supplementary Material online). Importantly, the distribution of divergence is qualitatively different across comparisons. Divergence is higher between VX34 and QX1410 than between N2 and XZ1516 across all chromosomes other than chromosome V, which harbors a notable excess of SNVs relative to the other autosomes in *C. elegans* (supplementary fig. S10, Supplementary Material online). We also find that the X chromosome divergence between QX1410 and VX34 is similar to that found on the autosomes (one SNV per 131 bp for the autosomes and one SNV per 147 bp for the X chromosome), in contrast to AF16 and QX1410 and both *C. elegans* comparisons, where SNP density is ~50% less on the X chromosomes relative to the autosomes (supplementary fig. S10, Supplementary Material online). Together, our results add to the growing number of studies showing that the genomes of selfing *Caenorhabditis* species harbor unexpectedly high levels of genetic diversity.
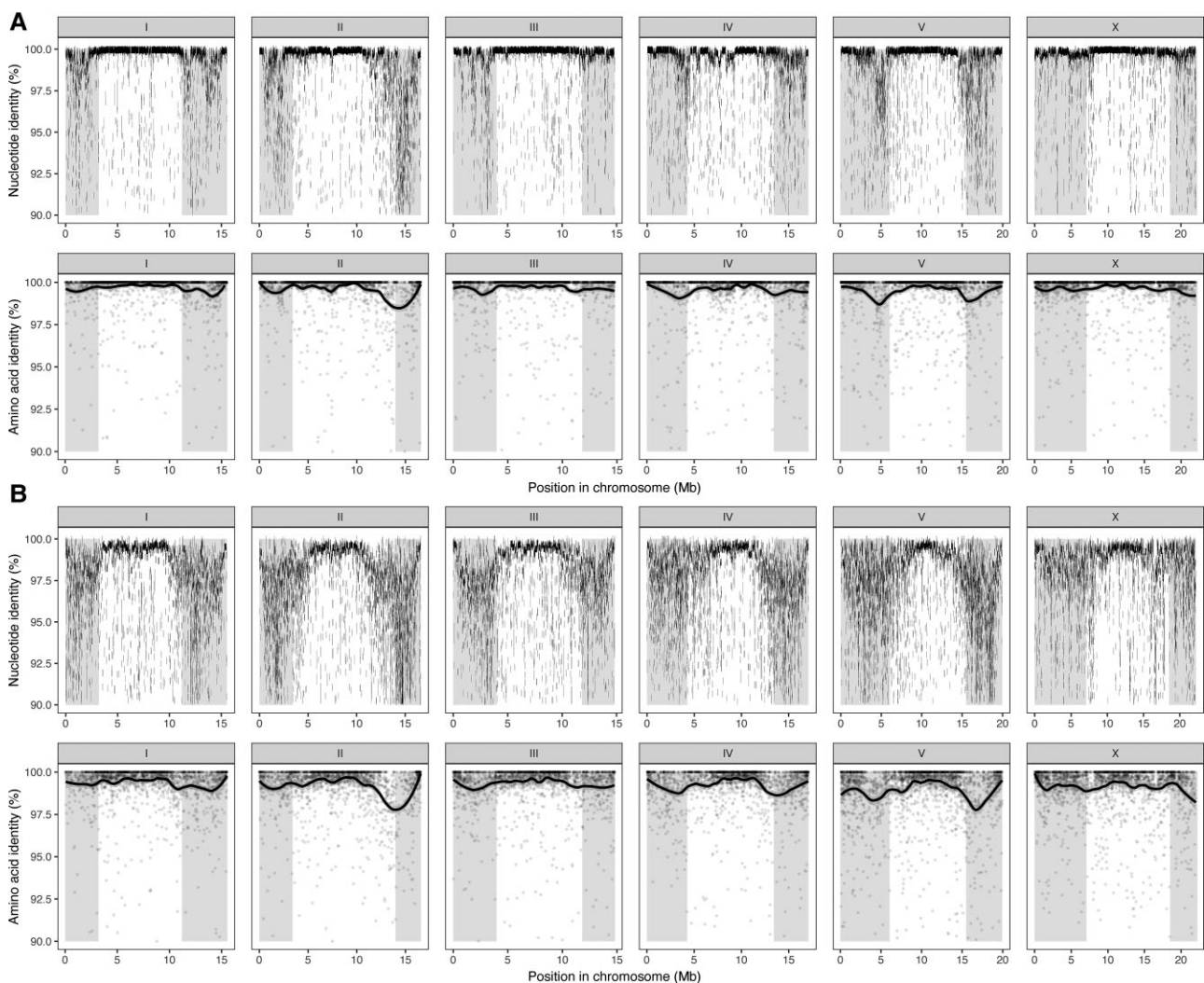
Fig. 3.—Genome-wide divergence between three *C. briggsae* strains. Genomes were aligned using nucmer and aligned regions of 1 kb or longer are shown. Conserved protein sequences were identified using OrthoFinder and aligned using MAFFT; lines represent LOESS smoothing curves fitted to the amino acid identity data. Grey shading indicates chromosome arm regions defined previously. (*A*) Nucleotide identity between aligned regions of QX1410 and AF16 genomes, and amino acid identity of protein sequences conserved between QX1410 and AF16. (*B*) Nucleotide identity between aligned regions of QX1410 and VX34 genomes, and amino acid identity of protein sequences conserved between QX1410 and VX34.

Selfing *Caenorhabditis* Species Have Undergone More Genome Rearrangements Than Their Outcrossing Sister Species

Early comparisons between the *C. elegans* and *C. briggsae* genomes revealed a strikingly high rate of intrachromosomal rearrangement when compared with other taxa such as *Drosophila* (Coghlan and Wolfe 2002; Hillier et al. 2007). Intriguingly, the rate of rearrangement was higher than expected from amino acid divergence alone: the rate of amino acid divergence was two times higher than in *Drosophila*, but the rate of rearrangement was four times higher (Coghlan and Wolfe 2002). One possible reason for this discrepancy is that both *C. elegans* and *C. briggsae* reproduce predominantly by self-fertilization. Evolutionary

theory predicts that rearrangements that are deleterious in heterozygotes are more likely to become fixed in populations of selfing species than they are in outcrossing species because selfers reach homozygosity more quickly (Lande 1979; Charlesworth 1992). Thus, over time, the genomes of selfing species are expected to undergo more rearrangement than their outcrossing relatives. In *Caenorhabditis*, selfing has evolved three times independently (in *C. elegans*, *C. briggsae*, and *C. tropicalis*; fig 4*A*; [Kiontke et al. 2004, 2011]) and chromosome-level reference genomes are now available for *C. inopinata* and *C. nigoni*, the outcrossing sister species of *C. elegans* and *C. briggsae*, respectively (Kanzaki et al. 2018; Yin et al. 2018). Although a chromosomally level reference genome was recently published for *C. tropicalis*
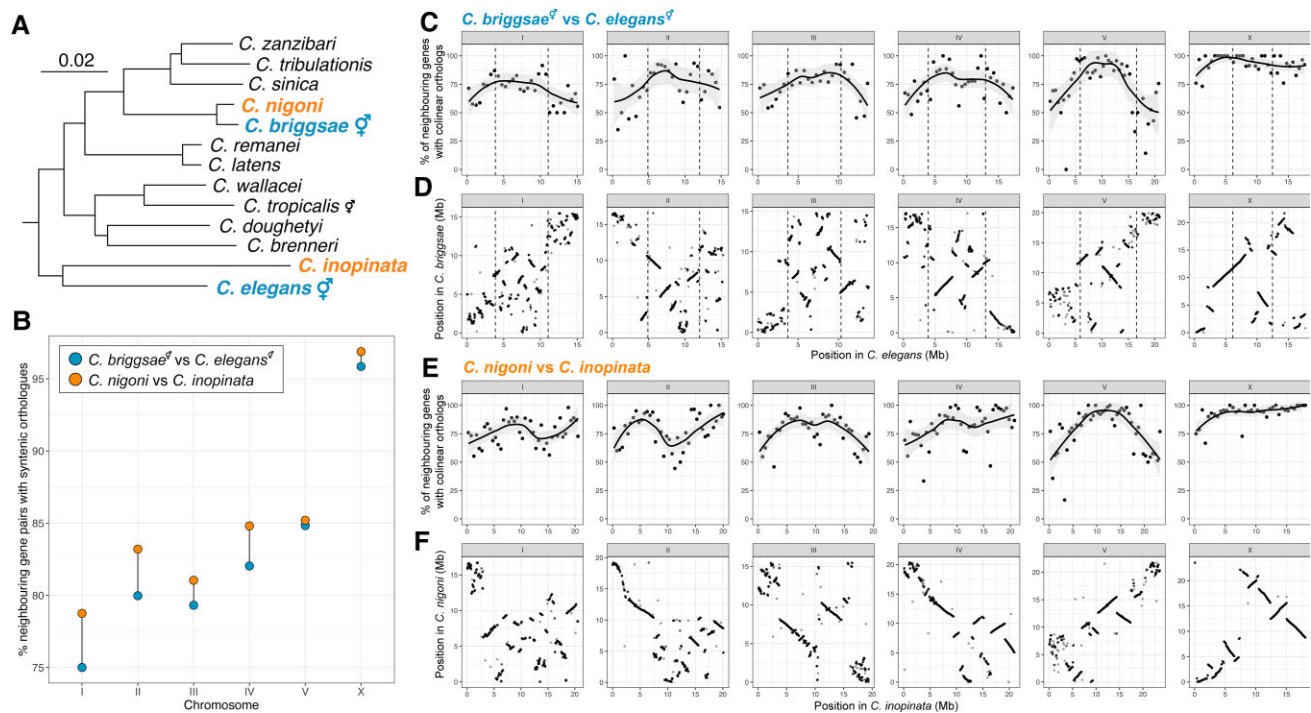
Fig. 4.—Selfing species have undergone more genome rearrangement than their outcrossing sister species. (*A*) *Caenorhabditis* phylogeny showing relationships within the *Elegans* group (Stevens et al. 2020). The species under comparison are highlighted in blue (selfers) or orange (outcrossers). Branch lengths are in substitutions per site; scale is shown. (*B*) Percentage of neighboring gene pairs in each chromosome with collinear orthologs between the two selfing and two outcrossing species. (*C*) The proportion of neighboring genes in 500 kb windows of the *C. elegans* genome that have collinear orthologs in the *C. briggsae* genome. Solid represent LOESS smoothing functions fitted to the data. Dotted lines represent the positions of the recombination rate domain boundaries ("arms" and "centers") in *C. elegans* (Rockman and Kruglyak 2009). (*D*) Positions of 9,395 one-to-one orthologs in the *C. elegans* and *C. briggsae* genomes. Dotted lines represent the positions of the recombination rate domain boundaries ("arms" and "centers") in *C. elegans*. (*E*) The proportion of neighboring genes in 500 kb windows of the *C. inopinata* genome that have collinear orthologs in the *C. nigoni* genome. Lines represent LOESS smoothing functions fitted to the data. (F) Positions of 9,395 one-to-one orthologs in the *C. inopinata* and *C. nigoni* genomes.

(Noble et al. 2021), no chromosomally level reference genome for its outcrossing sister species, *C. wallacei*, has been published.

We sought to determine if the unusually high intrachromosomal rearrangement rate previously observed in *C. elegans* and *C. briggsae* was related to their reproductive mode. As the rate of rearrangement is so high in the genomes of *Caenorhabditis* species, accurately inferring individual rearrangement events is challenging. To circumvent this problem, we identified single-copy orthologs across all four taxa and measured synteny between each genome by counting the number of neighboring gene pairs that had collinear orthologs. Consistent with previous work, we find that genomes of *C. elegans* and *C. briggsae* are highly rearranged relative to one another: an average of 17.1% of neighboring genes in *C. elegans* are noncollinear in *C. briggsae* (fig. 4B) and the autosomal arms are substantially more rearranged than the autosomal centers (fig. 4C and D). Moreover, the X chromosome is substantially more syntenic than the autosomes

(95.6% compared with a mean of 80.1%) with the three largest blocks of collinear genes (288, 185, 145) all found on the X chromosome (fig. 4C and D). Consistent with expectations from evolutionary theory, we find that the genomes of *C. elegans* and *C. briggsae* are more highly rearranged than their outcrossing sister species, *C. inopinata* and *C. nigoni* (17.1% of neighboring genes are rearranged in the selfers compared with 15.0% in the outcrossers; fig. 4E and F). This difference is despite the outcrossing pair of species being more distantly related than the corresponding pair of selfing species (an average of 0.11 amino acid substitutions per site between *C. elegans* and *C. briggsae* compared with 0.13 amino acid substitutions per site between *C. inopinata* and *C. nigoni*; fig. 4A). Although all six chromosomes are less syntenic in selfers than in outcrossers, the difference is not distributed equally. Chromosome I shows the largest reduction in synteny in selfers with ~3.8% more rearrangement between neighboring orthologs when compared with the outcrossers. By contrast, chromosome

V and the X chromosome, which is hemizygous in males, show comparatively small differences in synteny (0.3% and 1.0% more rearrangements between neighboring orthologs in selfers than in outcrossers, respectively).

However, we note that the patterns we observe could be explained by change or stasis in one of the four species, rather than a general increase in the rearrangement rate in selfing species. To address this point, we compared each species to an independent outcrossing species, *C. remanei*, for which a chromosome-level genome assembly was recently published (Teterina et al. 2020). Consistent with our previous results, we find that the genomes of *C. inopinata* and *C. nigoni* are more syntenic (86.4% and 87.8%, respectively) with *C. remanei* than their selfing sister species, *C. elegans* and *C. briggsae* (84.4%, and 87.1, respectively) (supplementary fig. S11, Supplementary Material online). Thus, it appears that the genomes of *C. briggsae* and *C. elegans* species have undergone a higher rate of rearrangement than their outcrossing relatives.

Discussion

High-Quality Reference Genomes for Two *C. briggsae* Strains

The AF16 reference genome was sequenced nearly 20 years ago using a combination of Sanger-based shotgun sequencing and a physical map generated by sequencing fosmid and BAC libraries (Stein et al. 2003). Since that time, advances in sequencing technologies, particularly long-read sequencing and high-throughput mapping approaches, have meant that it is now relatively straightforward to generate reference genomes that far surpass the quality of their predecessors. Despite being one of the highest quality reference genomes available for any eukaryotic organism, long-read sequencing of the *C. elegans* laboratory reference strain recently extended the reference genome by 1.8 Mb, largely by expanding previously collapsed repeated regions, including the rDNA cluster (Tyson et al. 2018; Yoshimura et al. 2019). Our new *C. briggsae* reference genomes, generated using Oxford Nanopore long-read and Hi-C data, are substantially more contiguous and complete than the existing AF16 reference genome. However, despite the quality of our raw data and the relatively small size of the *C. briggsae* genome, some regions of the QX1410 and VX34 genomes remain unresolved, particularly in the subtelomeric regions, which are known to be repetitive (Kim et al. 2019; Yoshimura et al. 2019). It is possible that the base-level accuracy of our long-read data, which we estimate to be ~94%, may have prevented large and highly similar repeats from being resolved during assembly. Technologies capable of producing long reads with high base-level accuracy, such as PacBio HiFi sequencing (Wenger et al. 2019), are now available and can lead

to substantial improvements in assembly contiguity, largely because of the ability to resolve large, near-identical repeats (Nurk et al. 2020). Future resequencing efforts using these technologies may fully resolve these complex regions of the *C. briggsae* genome.

Despite substantial differences in experimental design, our estimates of the recombination rate in the *C. briggsae* genome, and the physical position of the rate boundaries are largely congruent with previous estimates (Hillier et al. 2007; Ross et al. 2011). One region where we notice a large difference is the inferred physical position of the left arm-center boundary on chromosome II, which was estimated to be at ~3.4 Mb in our analysis but 4.53 Mb in the analysis of Ross *et al.* 2011. In our recombination map, the region surrounding the boundary shows a gentle change in recombination rate, leading to a 5–10-fold increase in standard error of the boundary position relative to other arm-center boundaries. With the exception of the left arm of I and right arms of II and III, our within-domain estimates of recombination rates are similar to previous measurements. The linear segment fitted to the left arm of chromosome I may be skewed because of the lack of a resolved subtelomeric region in the reference genome. In chromosome II, the disparity in marker density between the left and right arms prevented proper estimation of the left arm-center boundary. Subsequently, the skew in the fitted segment may explain the massive increase in recombination rate of the chromosome II right arm relative to other arm regions and to previous measurements.

Improved Approaches to Automated Protein-Coding Gene Prediction

Complete and accurate gene models are an essential resource to study the biology of any organism. Because of the complexity of protein-coding genes in eukaryotes, all currently available gene prediction tools fail to resolve the structure of most genes correctly (Mathé et al. 2002). For example, the widely used gene prediction pipeline BRAKER only predicts 55% of the *C. elegans* genes accurately (Hoff et al. 2016). Recent advances in long-read RNA-seq technologies now enable the sequencing of full-length transcripts that serve as an ideal template to accurately infer gene structures (Amarasinghe et al. 2020). Using the *C. elegans* N2 reference genes as a truth set, we developed a gene prediction pipeline that effectively combines long- and short-read data and leads to substantial improvements in the sensitivity and BUSCO completeness relative to gene sets predicted using either dataset alone. Based on protein sequence length similarity to the *C. elegans* proteome, we show that our QX1410 and VX34 gene models generated using this pipeline have improved accuracy relative to AF16 gene models generated using other automated approaches (WS255 release,

personal communication with WormBase staff). However, our benchmarks on *C. elegans* also reveal thousands of gene predictions that disagree with curated reference models, suggesting that prediction errors remain a common problem, even with long-read data. Manual inspection of gene models and underlying transcriptome data often reveals correctable mistakes such as the retention of noncoding sequences, incomplete coding sequences, missing or additional exons, and fused or split genes. Recent manual curation efforts in *Pristionchus pacificus* and *H. contortus* have led to substantial improvements in the quality of the reference annotation for these two species (Doyle et al. 2020; Rödelsperger 2021). Manual curation using multiple sources of experimental evidence will be necessary to further improve the quality of the QX1410 genome annotation.

Hyper-Divergent Regions Punctuate the Genomes of Wild *C. briggsae* Isolates

The availability of multiple reference genomes for *C. briggsae* allowed us to investigate the pattern of divergence across the genome. *Caenorhabditis briggsae* is known to harbor higher levels of genetic diversity than *C. elegans*, which appears to be explained by the existence of several, well defined phylogeographic groups (Cutter et al. 2006; Félix et al. 2013). Indeed, we found that the genome-wide divergence between QX1410, a member of the "Tropical" group, and VX34, a divergent strain isolated from China, was higher than that between the *C. elegans* laboratory strain N2 and the XZ1516, the most divergent *C. elegans* strain currently known (Lee et al. 2021). A previous population genomic study identified two *C. briggsae* strains isolated in Kerala, India (JU1341 and JU1348) that were far more divergent than any others, including VX34 (Thomas et al. 2015), and thus our data suggest that within-species divergence in *C. briggsae* is substantially higher than is currently known for *C. elegans*. Moreover, the X chromosome in the Keralan strains shows higher levels of divergence than the autosomes, a pattern that is mirrored in the outcrossing sister species, *C. nigoni* (Thomas et al. 2015). High-quality genomes for these strains may provide important insights into the mechanisms and patterns of genomic divergence that preceded speciation.

The presence of hyper-divergent haplotypes in *C. briggsae*, even between the "Tropical" QX1410 and AF16 strains, means that the regions are now known to exist in all three selfing *Caenorhabditis* species (Thompson et al. 2015; Lee et al. 2021; Noble et al. 2021). Although their origins remain unclear, the prevailing hypothesis is that these regions represent remnants of genetic diversity present in the outcrossing ancestor that have been maintained by long-term balancing selection since the evolution of selfing (Thompson et al. 2015; Lee et al. 2021). In *C. elegans*, these haplotypes are enriched for genes involved in environmental responses and include the large chemosensory G-protein-coupled receptors and nuclear hormone receptors (Thompson et al. 2015; Lee et al. 2021). Intriguingly, the structure of the *C. elegans* genome, with large rapidly evolving gene families being enriched on autosomal arms, is known to be conserved in *C. briggsae* (Hillier et al. 2007) raising the possibility that balancing selection has acted on loci with similar functions, or even orthologs, in both species. Furthermore, the presence of these haplotypes in *C. briggsae* also provides an opportunity to conclusively identify their origins. Unlike in *C. elegans*, which lacks a closely related sister species (Kanzaki et al. 2018), *C. briggsae* is estimated to have diverged from its outcrossing sister species, *C. nigoni*, around 3.5 million years ago (Thomas et al. 2015). Future species-wide genome sequencing efforts in *C. briggsae* and *C. nigoni*, especially those involving high-quality genomes of wild isolates generated using long-read data, could identify whether hyper-divergent haplotypes are shared between these two species and whether their divergence is consistent with long-term balancing selection or recent adaptive introgression.

Selfing Species Have Undergone More Genome Rearrangement Than Their Outcrossing Sister Species

The transition from outcrossing to self-fertilization has a profound impact on the evolutionary forces that shape the genome. The predicted genomic consequences of selfing, which are collectively known as the "genomic selfing syndrome" (Cutter 2019), include an increased rate of genomic rearrangement, as structural rearrangements that are deleterious when heterozygous will be more likely to become fixed in highly selfing populations because they reach homozygosity more quickly (Lande 1979; Charlesworth 1992). One of the first comparative genomic analyses performed between *C. elegans* and *C. briggsae*, two species that have independently evolved self-fertilization from an outcrossing ancestor, was to compare synteny, or gene order, between their genomes, which revealed an unusually high rate of intrachromosomal rearrangement (Coghlan and Wolfe 2002; Hillier et al. 2007). Using chromosome-level reference genomes for five *Caenorhabditis* species (*C. elegans* Sequencing Consortium 1998; Kanzaki et al. 2018; Yin et al. 2018; Teterina et al. 2020), we have shown that, consistent with theoretical predictions, the genomes of *C. elegans* and *C. briggsae* have more highly rearranged genomes than their outcrossing sister species, *C. inopinata* and *C. nigoni*. The pattern of reduced synteny in selfers is true across all chromosomes but is less pronounced on chromosomes V and X. A potential explanation for the X chromosome difference is that, in an outcrossing *Caenorhabditis* species, the X is only heterozygous in females (males are hemizygous for X), meaning the effect of selfing on the rearrangement rate in the X chromosome

would be expected to be half of that observed in autosomes (assuming a 50/50 ratio of males and females). However, the X chromosome shows a far reduced rearrangement rate than the autosomes in both outcrossers and selfers, suggesting gene order on the X is under selective constraints that are unrelated to reproductive mode. We also note that the overall decrease in degree of synteny seen in selfing genomes is relatively small, which is expected based on the relatively short time each species is believed to have been reproducing via self-fertilization (within the last 3.5 and 4 million years for *C. briggsae* and *C. elegans*, respectively [Cutter et al. 2008; Thomas et al. 2015]). Therefore, although the rates inferred by Coghlan and Wolfe (2002) were likely biased by comparing two selfing species, it remains true that *Caenorhabditis* genomes undergo a high rate of intrachromosomal rearrangement. Because we only surveyed two selfing and outcrossing species pairs, it remains possible that the differences we observed are the result of coincidental lineage-specific variation in rearrangement rate. Therefore, it will be interesting to know if the pattern holds true in the remaining selfing and outcrossing *Caenorhabditis* species pair, *C. tropicalis* and *C. wallacei*, and in related genera that have independently evolved self-fertilization, such as *Oscheius* (Baïlle et al. 2008).

Future Outlook

Although an improved reference genome for *C. briggsae* is an essential step in our efforts to understand genome evolution in *Caenorhabditis*, it is only a beginning. In recent years, an improved understanding of the natural ecology of *Caenorhabditis* nematodes has led to a dramatic increase in the discovery of new species with over 60 species now in laboratory culture, and multiple isolates are available for many species (Kiontke et al. 2011; Félix et al. 2014; Ferrari et al. 2017; Kanzaki et al. 2018; Crombie et al. 2019; Stevens et al. 2019; Dayi et al. 2021). Although some species have high-quality reference genomes (*C. elegans* Sequencing Consortium 1998; Kanzaki et al. 2018; Yin et al. 2018; Teterina et al. 2020; Noble et al. 2021), the majority has been sequenced using short-reads only and therefore have highly fragmented genome assemblies that obscure the higher-order structure of the genome and complicate downstream analyses (Stevens et al. 2019). Efforts to generate chromosome-level reference genomes from across the *Caenorhabditis* phylogeny and in related genera will help to answer several long-standing questions about *Caenorhabditis* and nematode genome evolution. Why do genes rarely move between chromosomes despite strikingly high rates of within-chromosomal rearrangement? What is the origin of the "arms" and "centers" recombination landscape, and how well conserved is it? And why, despite nematode

chromosomes being holocentric, is the *C. elegans* karyotype so highly conserved in related species? Moreover, in the majority of cases, only a single strain has been sequenced for each species, and studies of within-species genetic variation have been largely restricted to selfing species. Resequencing datasets, particularly those involving long-read data, will reveal the distribution and levels of genetic diversity in outcrossing species and provide an important context to the recent discoveries of hyper-divergent haplotypes in selfers. These new insights, made possible by advances in sequencing technology, an improved understanding of ecology, and intense sampling efforts, will help to place *C. elegans* and the vast body of knowledge of its biology within a rich evolutionary context.

Methods

Nematode Culture

Nematodes were reared at 20°C using *Escherichia coli* OP50 bacteria grown on a modified nematode growth medium (NGMA), containing 1% agar and 0.7% agarose to prevent animals from burrowing (Andersen et al. 2014).

Short-Read Illumina Sequencing

To extract DNA, we transferred nematodes from three recently starved 10 cm NGMA plates 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 Blood and Tissue DNA isolation kit (cat# 69506, QIAGEN, Valencia, CA, USA) following established protocols (Cook et al. 2016). The DNA concentration was determined for each sample using the Qubit dsDNA Broad Range Assay Kit (cat# Q32850, Invitrogen, Carlsbad, CA, USA). For high-coverage sequencing, libraries were generated with New England BioLabs NEBNext® Ultra™ II FS DNA Library Prep (NEB, Ipswich, MA, USA). Samples were sequenced at the Duke Center for Genomic and Computational Biology, Novogene, or the Northwestern Sequencing facility, NUSeq. All samples were sequenced on the NovaSeq 6000 platform (paired-end 150 bp reads).

For low-coverage sequencing, libraries were generated using a modified Illumina Nextera Sample Prep (Illumina, FC-121-1030) protocol. For each sample, 0.16 ng of DNA was tagmented for 5 min at 55 °C with 2.5 µl of a 1/35 dilution of the Illumina Transposome in a tris buffer (10 mM Tris–HCl, pH 8.0; 5 mM MgCl$_2$). Tagmented samples were then amplified and barcoded using ExTaq (TaKaRa,

RR001B) and custom primers. Resulting libraries were pooled and sequenced on the Illumina MiSeq.

Long-Read Oxford Nanopore Sequencing

Nematodes were collected using the same technique as for short-read sequencing but on 14 10 cm plates instead of three. Animals were transferred from plates using 25 ml of M9 and washed into a 50 ml conical. The 300–500 μl worm pellets were submitted to the DNA Technologies and Expression Analysis Cores at University of California, Davis for High Molecular Weight gDNA extraction, library preparation, and sequencing on the Oxford Nanopore PromethION system.

Hi-C Library Preparation

The Hi-C libraries were prepared using a modified protocol based on a method described previously (Crane et al. 2015). Briefly, ~12,000 adult nematodes were harvested and washed in M9 buffer. The animals were crosslinked with 2% (v/v) formaldehyde solution, then dounced to disrupt pellets in 1 ml lysis buffer (10 mM Tris–HCl, pH = 8.0, 10 mM NaCl and 0.1% [v/v] protease inhibitors). The chromatin was digested overnight by DpnII, then incubated at 65 °C for 15 min to deactivate the enzyme. The DNA ends were biotinylated at 23 °C for 4 h and blunt-end ligated with T4 DNA ligase at 16 °C for 4 h. Proteinase K (50 μl of 10 mg/ml) was added to each tube to reverse crosslinks and degrade proteins. Then, equal volumes of phenol and chloroform (1:1) were added per tube and DNA purified using 15 ml phase lock tubes (1500 g for 5 min). Then, the aqueous phase was transferred to the clear 35 ml tube and precipitated with 10% volume of 3 M sodium acetate and 2.5 volumes of ice cold 100% ethanol. The pellet was dried at room temperature and then dissolved in 5 ml of TE buffer. Next, biotin was removed from the unligated ends by adding 5 μl of T4 DNA polymerase (NEB) to 5 μg of DNA, before shearing the DNA to a size of 100–300 bp using the Covaris M220 apparatus. Biotinylated fragments were pulled down using streptavidin beads and resuspended in a ligation buffer. Then, Illumina indices were added along with adapters. The beads were pelleted using a Magnetic Particle Concentrator (Thermofisher) for 1 min, and washed several times before being resuspended the beads in 20 μl NEBuffer 2 (New England Biolabs). The final library was generated using the Illumina TruSeq kit (Illumina) and sequenced using the Illumina HiSeq 4000 platform, yielding 50 bp paired-end reads.

RNA Extraction

Three sets of samples were collected for *C. briggsae* strains QX1410 and VX34 and *C. elegans* PD1074 in order to maximize transcript representation. For each strain, we sampled a mixed-staged population prepared by chunking plates

every 2 days for several generations, a male-enriched population created by setting up crosses and expanding the population for 2–3 generations, and a starved population by allowing plates to starve out so that they contained dauers and arrested L1 and L2 larvae. For these cultures, nematodes were reared as elsewhere, except that standard NGMA was used (no agarose). Each sample was collected from one 10 cm plate in 100 ml S Basal, flash frozen in liquid nitrogen, and stored at −80 °C. RNA was extracted using 1 ml of the TRIzol reagent (Invitrogen, catalog no. 15596026) following the manufacturer's protocol except that 1 ml of acid-washed sand (Sigma, catalog no. 27439) was added to aid homogenization. RNA was resuspended in nuclease-free water. A Nanodrop spectrophotometer (ThermoFisher) was used to assess the purity of the extracted RNA, a Bioanalyzer (Agilent) was used to determine RNA integrity with the RNA 6000 Pico Kit (Agilent, catalog no. 5067-1513), and a Qubit (ThermoFisher) was used to determine RNA concentration with the Qubit RNA HS Assay kit (ThermoFisher, catalog no. Q32852). Following QC, we pooled 1.5 mg RNA from each sample (mixed-stage, male-enriched, and starved) and used the RNeasy MinElute Cleanup kit (Qiagen, catalog no. 74204) to further purify and concentrate the pooled RNA. We eluted the RNA in nuclease-free water and performed another round of QC as before.

Long-Read RNA-seq

300 ng total RNA was used to prepare each PacBio Iso-Seq full-length transcript sequencing library. Libraries were prepared in the Duke Center for Genomic and Computational Biology's Sequencing and Genomic Technologies Core Facility using the NEBNext Single Cell/Low Input cDNA Synthesis and Amplification Module (NEB, catalog no. E6421) and SMRTbell Express Template Prep Kit 2.0 (Pacific Biosciences, catalog no. 100-938-900). Each library was sequenced with three SMRT cells.

Short-Read RNA-seq

We prepared Illumina RNA-seq libraries of *C. briggsae* strains QX1410 and VX34 in one 96-well plate simultaneously. For each sample, the NEBNext Poly(A) mRNA Magnetic Isolation Module (New England Biolabs, catalog no. E7490L) was used to purify and enrich mRNA from 1 μg of total RNA. We performed RNA fragmentation, first and second strand cDNA synthesis, and end-repair processing using the NEBNext Ultra II RNA Library Prep with Sample Purification Beads (New England Biolabs, catalog no. E7775L). Adapters and unique dual indexes in the NEBNext Multiplex Oligos for Illumina (New England Biolabs, catalog no. E6440L) were used to adapter-ligate the cDNA libraries. We performed all procedures according to manufacturer protocols. We determined the concentration of each

RNA-seq library using Qubit dsDNA BR Assay Kit (Invitrogen, catalog no. Q32853). RNA-seq libraries were pooled and qualified with the 2100 Bioanalyzer (Agilent) at Novogene, CA, USA. The pooled libraries were sequenced on a single lane of an Illumina NovaSeq 6000 platform, yielding 150-bp paired-end (PE150) reads.

RIL Construction, Genotyping, and Cross-Object Creation

RILs between QX1410 and VX34 were constructed by first generating heterozygous individuals by crossing males of each parent strain to hermaphrodites of the other strain. These heterozygous individuals had maternal contributions (mitochondria) from either parent, so four different genotypes were generated: males and hermaphrodites each with either QX1410 or VX34 mitochondria. Heterozygous males from each parental cross were crossed to both types of heterozygous hermaphrodites in four different crosses. Twenty-five hermaphrodite cross progeny from each of these four crosses were picked to individual plates for a total of 100 independent recombinant progeny. These individuals were selfed by single-animal passage for ten generations. After which time, each RIL was cryopreserved and its genome was sequenced.

Once sequenced, raw reads for the 99 lines were processed for genotyping using the Andersen Lab's nil-ril nextflow pipeline (https://github.com/AndersenLab/nil-ril-nf). The raw reads were aligned to the QX1410 reference genome. Strains that were run multiple times were merged into a single BAM file. Variants were called and put into a dataset along with the parental genotypes. A hidden-markov-model (HMM) was used to fill in missing genotypes from the low-coverage data. The HMM VCF was then used to generate a genotype coordinate flat file for the position of each variant and the parental genotype of each strain at each position.

Genome Assembly

We downsampled the ONT data for QX1410 and VX34 to ~200× coverage using FiltLong (v0.2.0; https://github.com/rrwick/Filtlong), based on a genome size of 106 Mb. We assembled the subsampled long reads independently for each strain using Canu r10117 (Koren et al. 2017), Flye v2.8.1-b1676 (Kolmogorov et al. 2018) and wtdbg2 v0.0 (Ruan and Li 2019) using default parameters. We used nucmer v3.1 (Delcher et al. 2003) to align each assembly to the current version of the AF16 genome. The Flye assemblies were consistently more contiguous than the Canu and wtdbg2 assemblies. We aligned the ONT reads to the Flye assemblies using minimap2 v2.17-r941 (Li 2018) and provided the resulting alignments to racon v1.4.13 (Vaser et al. 2017) to perform error correction using the parameters recommended by ONT (-m 8 -x -6 -g -8 -w 500).

We then provided these corrected assemblies and the ONT reads to Medaka v1.1.2 (https://github.com/nanoporetech/medaka) to correct sequencing errors. We corrected remaining sequencing errors by aligning a paired-end Illumina dataset to the assemblies using bwa mem v0.7.17-r1188 (Li 2013) and providing the resulting alignments to Pilon v1.23 (Walker et al. 2014) (using the –fix bases parameter).

Hi-C Scaffolding

To scaffold the polished assemblies into complete chromosomes, we downsampled the Hi-C data to ~50× coverage using seqtk v1.3-r106 (available at https://github.com/lh3/seqtk). We used the Juicer/3D-DNA pipeline v1.6 and v180114 (Durand et al. 2016; Dudchenko et al. 2017) to align the Hi-C data to the assembly and scaffold into chromosomal scaffolds (using default parameters, with "*Dpn*II" as the restriction enzyme). We noticed that 3D-DNA was erroneously breaking contigs in highly repetitive regions, and that this behavior could not be suppressed by turning off misjoin correction (-r 0). To avoid this, we used the early exit mode in 3D-DNA (–early-exit) and then manually edited the assembly file to create large chromosomal scaffolds based on the corresponding Hi-C contact map. The resulting assemblies both comprised six chromosomal scaffolds. The QX1410 genome had five unplaced scaffolds (15–71; 183 kb in total span) and the VX34 genome had three unplaced contigs (37–66; 141 kb in total span).

Mitochondrial Genome

We identified the mitochondrial contig by searching the genome assembly for hits to the AF16 mitochondrial genome with BLASTN (Camacho et al. 2009). As is common for circular genomes, the assembled contig consisted of a "concatemer" composed of more than one copy. To resolve this step, we used mitoHiFi (v1; available at https://github.com/marcelauliano/MitoHiFi) to decircularize the mitochondrial genome.

Curation and QC

We assessed the base-level accuracy of the Hi-C-scaffolded assembly by estimating QV score using Merqury v1.1 (Rhie et al. 2020) and the Illumina short-read library. For the unplaced contigs, we manually inspected each contig and the corresponding reads using gap5 (Bonfield and Whitwham 2010). We also aligned each contig to the assembly using BLASTN. We found that all unplaced contigs comprised redundant repeats, several of which were low complexity (including a contig that was composed entirely of repeat units of the telomeric hexamer, TTAGGC). All unplaced contigs were removed from the final assembly. The resulting, curated chromosomes were reoriented relative to the *C. briggsae* AF16 reference genome.

Repeat Masking

Prior to gene prediction, we masked repetitive sequences using a custom repeat library. The approach used to generate custom repeat libraries for nematode genomes has been previously described (Berriman et al. 2018; Teterina et al. 2020). In summary, we first identified repetitive sequences de novo using RepeatModeler from RepeatMasker v2.0.1 (Smit et al. 2015). We identified transposable elements using TransposonPSI (available at http://transposonpsi. sourceforge.net/). We identified long terminal repeat (LTR) retrotransposons using LTRharvest and LTRdigest from Genome Tools v1.6.1 (Ellinghaus et al. 2008; Gremme et al. 2013). Putative LTR retrotransposon sequences were identified with the *gt-ltrharvest* tool, followed by annotation with the *gt-ltrdigest* tool using HMM profiles from Gypsy Database v2.0 (Llorens et al. 2011), and Pfam domains (Finn et al. 2014) selected from supplementary tables SB1 and SB2, Supplementary Material online of Steinbiss et al. (2009). Sequences from *gt-ltrdigest* were filtered using the *gt-select* tool to remove sequences without conserved protein domains. Additionally, we retrieved all Rhabditid repeats from RepBase (Bao et al. 2015) and Dfam (Hubley et al. 2016), respectively. Newly generated and retrieved repeat libraries were merged into a single redundant repeat library. Repeats in the merged library were clustered using VSEARCH v2.14.2 (Rognes et al. 2016) and classified with the RepeatClassifier tool from RepeatModeler. We removed unclassified repeats that had BLASTX hits to *C. elegans* proteins and soft-masked the genome assemblies using RepeatMasker.

Protein-Coding Gene Prediction

We generated protein-coding gene predictions using BRAKER v2.1.6 (Hoff et al. 2019). In summary, we aligned short RNA-seq reads for each strain to their respective soft-masked genome using STAR v2.7.3a (Dobin et al. 2013) in two-pass mode with a maximum intron size of 10 kb. We then supplied sequence alignments and soft-masked genome assemblies to the BRAKER pipeline. Additionally, we generated high-quality transcripts from Pacific Biosciences (PacBio) long RNA reads using the IsoSeq3 pipeline v3.4.0 (available at https://github.com/PacificBiosciences/IsoSeq). We aligned PacBio high-quality transcripts for each strain to their respective genome using minimap2 (Li 2018). We supplied long-read transcript alignments to StringTie v2.1.2 (Kovaka et al. 2019) and performed transcriptome assembly. The coding sequences (CDS) of the assembled transcripts were predicted using Transdecoder v5.5.0 (available at https://github.com/TransDecoder/TransDecoder). We extracted StringTie gene models with incomplete CDS using *agat_sp_remove_incomplete_gene_models.pl* script from AGAT v0.8.1 (Dainat et al. 2022). We repaired the CDS of most incomplete StringTie models using

agat_sp_fix_longest_ORF.pl from AGAT. We removed any modified-CDS StringTie models that remained incomplete using *agat_sp_remove_incomplete_gene_models* script from AGAT. We merged complete StringTie models with modified-CDS StringTie models using *agat_sp_merge_annotations.pl* from AGAT. We removed any StringTie gene models that had more than one noncoding exon, producing a final set of StringTie gene models. We then merged the final set of StringTie models with BRAKER gene models using *agat_sp_merge_annotations.pl* from AGAT. We fused genes that had transcripts with overlapping CDS using *agat_sp_fix_overlapping_genes.pl* from AGAT. Lastly, since AGAT only repairs overlaps when genes share a CDS, we identified genes that were fully overlapped by other genes and lacked a common CDS. We extracted the coordinates of the overlaps and removed genes within the coordinates using *agat_sp_filter_records_by_coordinate.pl* from AGAT. We then extracted the original BRAKER genes that were predicted in those regions and filled the empty coordinates using *agat_sp_merge_annotations.pl* from AGAT. For every gene, we removed redundant isoforms with identical CDS and intron chains. We renamed all annotated features IDs with cohesive prefixes using *agat_sp_manage_IDs.pl* from AGAT.

Gene Prediction Pipeline Benchmarks

We benchmarked the quality of our gene prediction pipeline using the *C. elegans* N2 reference annotation from WormBase (WS279) as a truth set. We generated gene models for the *C. elegans* reference genome using the gene prediction pipeline described previously. For short-read-based gene prediction, we generated an N2 mixed-population RNA-seq sample by pooling Illumina data from individual samples representing all larval stages and adults (retrieved from SRR953117, SRR953118, SRR953119, SRR953120, and SRR953121). To match the depth of our *C. briggsae* RNA-seq samples, we randomly downsampled the pooled N2 Illumina reads to 35 million reads. We compared individual BRAKER and StringTie models and merged gene models against the existing *C. elegans* reference annotation using GffCompare from GffRead (Pertea and Pertea 2020). We extracted sensitivity and precision statistics for base, exon, intron, intron chain, transcript, and locus level. We assessed the biological completeness of the N2 predictions using BUSCO v4.0.5 (Simão et al. 2015) with the Nematoda (odb10) database.

C. briggsae Gene Predictions

We generated gene models for *C. briggsae* QX1410 and VX34 using the gene prediction pipeline described above. To assess the quality of the *C. briggsae* gene predictions, we developed a script to search for reciprocal BLAST hits between protein sequences of *C. briggsae* predictions and the *C. elegans* reference annotation. We extracted protein

sequences from each gene prediction set using GffRead. For each protein sequence, we kept the best BLAST hit based on expected value and bit score. Two or more best hits were kept if their expected values and bit score were identical. We considered a pair of protein sequences to be reciprocal if the best hit of the first sequence matched the best hit of the second, and vice versa. Pair of sequences with multiple best hits were considered reciprocal if any of their hits were reciprocal. For each set of gene predictions, we counted the number of genes that have at least one reciprocal protein sequence (total matches). We compared the protein lengths of every predicted protein sequence against its reciprocal. We counted the number of genes that had the same protein length (1:1 hits) and the number of genes that were within 5% of the protein length (5% off hits) of its reciprocal. We assessed biological completeness using BUSCO with the Nematoda (odb10) database.

Genetic Map Construction and Domain Analysis

We filtered markers and estimated genetic distances using the *R/qtl* package (Broman et al. 2003). In summary, we read RIL genotypes into a cross-object using the *read.cross()* function with the "riself" parameter. We identified and removed 25 markers with distorted segregation patterns using the *geno.table()* function, and three markers that fell outside six major linkage groups using the *formLinkageGroups()* function. Additionally, we developed an algorithm to search and remove individual or groups of markers with local deviations in allele frequency equal or >4% relative to its neighboring markers. We removed 103 markers with deviations in allele frequency. We estimated genetic distances using *est.rf()* with the default Haldane map function. We manually removed 22 additional markers with exceedingly high recombination rate in close physical distance and reestimated genetic distances. Chromosomal domain analyses were performed using the *R/segmented* package (Muggeo 2003). Arm-center domain boundaries were defined using segmented linear regressions with two expected breakpoints. Because markers in the left arm of chromosome II were sparse, the left arm-center boundary was not defined as a segment breakpoint when including all markers in the initial linear fit. To estimate this missing arm-center boundary, we excluded markers after 15 Mb in the right end of chromosome II prior to linear fit. Tip domains were manually called by identifying the first pair of markers that showed a substantial change in recombination rate.

Assessing Genome-Wide Divergence Between the Three *C. briggsae* Genomes

To assess nucleotide divergence between the three *C. briggsae* strains, we used nucmer 3.1 (Delcher et al. 2003) to align VX34 and AF16 to the QX1410 reference genome. To measure amino acid divergence, we employed

an orthology clustering approach. Briefly, we identified and selected the longest isoform of each protein-coding gene for QX1410, AF16, and VX34 using AGAT v0.8.1 (Dainat et al. 2022). The isoform-filtered proteomes were then clustered into orthologous groups using OrthoFinder v2.5.2 (Emms and Kelly 2019). We selected all single-copy orthologous groups and aligned the protein sequenced with MAFFT v7.475 (Katoh and Standley 2013) and calculated the average amino acid identity using a custom script (available at https://github.com/AndersenLab/briggsae_reference_genome_MS). We also called SNVs using an assembly based approach. We aligned AF16 and VX34 to the QX1410 reference genome using minimap2 2.17-r941 (Li 2018) and provided the resulting PAF file to paftools v2.18-r1015 (available at https://github.com/lh3/minimap2) to call variants (setting the minimum alignment length to compute coverage and call variants to 1000). The resulting VCF file was filtered to contain only biallelic SNVs using bcftools v1.13 (Danecek et al. 2014) and SNV densities per chromosomes were calculated using bedtools v2.30.0 (Quinlan and Hall 2010).

Comparing Synteny Between Selfing and Outcrossing Species

We downloaded the genomes and annotation files for *C. elegans* (*C. elegans* Sequencing Consortium 1998), *C. inopinata* (Kanzaki et al. 2018), and *C. nigoni* (Yin et al. 2018) from WormBase (WS279) (Harris et al. 2020), and for *C. remanei* from NCBI (GCA_010183535.1) (Teterina et al. 2020) and used AGAT v0.8.1 (available at https://github.com/NBISweden/AGAT) (Dainat et al. 2022) to extract the longest isoform from each protein-coding gene. We generated two orthology clustering sets by clustering the isoform-filtered protein sequences with OrthoFinder v2.5.4 (Emms and Kelly 2019): one containing only the selfing/outcrossing species pairs (*C. elegans*, *C. inopinata*, *C. briggsae*, and *C. nigoni*) and a second that also included *C. remanei*. For both datasets, we selected orthogroups containing protein sequences that were single-copy across all species and extracted their corresponding coordinates using custom scripts (available at https://github.com/AndersenLab/briggsae_reference_genome_MS). To count the proportion of neighboring genes that had collinear orthologs, we assigned a number of each orthologous gene, corresponding to its order along each chromosome, and for each neighboring gene pair determined if their orthologs were collinear (i.e., had a distance of one). These data were then used to compute values for nonoverlapping 500 kb windows, identify blocks of collinear genes, and average synteny for each chromosome. We performed two comparisons (1) a direct comparison between *C. elegans* and *C. briggsae* (selfers), and between *C. inopinata* and *C. nigoni* (outcrossers), and

(2) a comparison between all four focal species and *C. remanei*. To compare our measures of synteny with amino acid divergence, we used ETE3 (Huerta-Cepas et al. 2016) to extract branch lengths (in amino acids substitutions per site) from a recently published *Caenorhabditis* phylogeny based on the protein sequences of 1,167 single-copy orthologs (Stevens et al. 2020).

Supplementary Material

Supplementary data are available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

Acknowledgments

Data Availability

Raw sequencing data and genome assembly and annotation files have been archived under the NCBI study accession PRJNA784955. Code associated with the analyses and figures can be found at https://github.com/AndersenLab/briggsae_reference_genome_MS.

Literature Cited

C. elegans Sequencing Consortium. 1998. Genome sequence of the nematode *C. elegans*: a platform for investigating biology. Science 282:2012–2018.

Amarasinghe SL, et al. 2020. Opportunities and challenges in long-read sequencing data analysis. Genome Biol. 21:30.

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

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

Baïlle D, Barrière A, Félix M-A. 2008. *Oscheius tipulae*, a widespread hermaphroditic soil nematode, displays a higher genetic diversity and geographical structure than *Caenorhabditis elegans*. Mol Ecol. 17:1523–1534.

Bao W, Kojima KK, Kohany O. 2015. Repbase update, a database of repetitive elements in eukaryotic genomes. Mob DNA. 6:11.

Barrière A, Félix M-A. 2005. High local genetic diversity and low outcrossing rate in *Caenorhabditis elegans* natural populations. Curr Biol. 15:1176–1184.

Ben-David E, et al. 2021. Ubiquitous selfish toxin-antidote elements in *Caenorhabditis* species. Curr Biol. 31:990–1001.

Berriman M, Coghlan A, Tsai IJ. 2018. Creation of a comprehensive repeat library for a newly sequenced parasitic worm genome. Protocol Exchan. doi:10.1038/protex.2018.054.

Bonfield JK, Whitwham A. 2010. Gap5—editing the billion fragment sequence assembly. Bioinformatics 26:1699–1703.

Broman KW, Wu H, Sen S, Churchill GA. 2003. R/qtl: QTL mapping in experimental crosses. Bioinformatics 19:889–890.

Camacho C, et al. 2009. BLAST+: architecture and applications. BMC Bioinform. 10:421.

Charlesworth B. 1992. Evolutionary rates in partially self-fertilizing species. Am Nat. 140:126–148.

Cho S, Jin S-W, Cohen A, Ellis RE. 2004. A phylogeny of *Caenorhabditis* reveals frequent loss of introns during nematode evolution. Genome Res. 14:1207–1220.

Coghlan A, Wolfe KH. 2002. Fourfold faster rate of genome rearrangement in nematodes than in *Drosophila*. Genome Res. 12:857–867.

Cook DE, et al. 2016. The genetic basis of natural variation in *Caenorhabditis elegans* telomere length. Genetics 204:371–383.

Crane E, et al. 2015. Condensin-driven remodelling of X chromosome topology during dosage compensation. Nature 523:240–244.

Crombie TA, et al. 2019. Deep sampling of Hawaiian *Caenorhabditis elegans* reveals high genetic diversity and admixture with global populations. Elife 8:e50465.

Cutter AD. 2019. Reproductive transitions in plants and animals: selfing syndrome, sexual selection and speciation. New Phytol. 224: 1080–1094.

Cutter AD, Félix M-A, Barrière A, Charlesworth D. 2006. Patterns of nucleotide polymorphism distinguish temperate and tropical wild isolates of *Caenorhabditis briggsae*. Genetics 173: 2021–2031.

Cutter AD, Wasmuth JD, Washington NL. 2008. Patterns of molecular evolution in *Caenorhabditis* preclude ancient origins of selfing. Genetics 178:2093–2104.

Dainat J, Hereñú D, LucileSol, pascal-git. 2022. *NBISweden/AGAT: AGAT-v0.8.1*. doi: 10.5281/zenodo.5834795.

Danecek P, Schiffels S, Durbin R. 2014. Multiallelic calling model in bcftools (-m). https://samtools.github.io/bcftools/call-m.pdf

Dayi M, et al. 2021. Additional description and genome analyses of *Caenorhabditis auriculariae* representing the basal lineage of genus Caenorhabditis. Sci Rep. 11:6720.

Delcher AL, Salzberg SL, Phillippy AM. 2003. Using MUMmer to identify similar regions in large sequence sets. Curr Protoc Bioinform. Chapter 10:Unit 10.3.

Dobin A, et al. 2013. STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29:15–21.

Dolgin ES, Charlesworth B, Baird SE, Cutter AD. 2007. Inbreeding and outbreeding depression in *Caenorhabditis* nematodes. Evolution 61:1339–1352.

Doyle SR, et al. 2020. Genomic and transcriptomic variation defines the chromosome-scale assembly of *Haemonchus contortus*, a model gastrointestinal worm. Commun Biol. 3:656.

Dudchenko O, et al. 2017. De novo assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. Science 356:92–95.

Durand NC, et al. 2016. Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. Cell Syst. 3:95–98.

Edgar RS, Wood WB. 1977. The nematode *Caenorhabditis elegans*: a new organism for intensive biological study. Science 198: 1285–1286.

Ellinghaus D, Kurtz S, Willhoeft U. 2008. LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. BMC Bioinform. 9:18.

Ellis RE, Sulston JE, Coulson AR. 1986. The rDNA of *C. elegans*: sequence and structure. Nucleic Acids Res. 14:2345–2364.

Emms DM, Kelly S. 2019. OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biol. 20:238.

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

Félix M-A, Braendle C, Cutter AD. 2014. A streamlined system for species diagnosis in *Caenorhabditis* (Nematoda: Rhabditidae) with name designations for 15 distinct biological species. PLoS One 9: e94723.

Félix M-A, Duveau F. 2012. Population dynamics and habitat sharing of natural populations of *Caenorhabditis elegans* and *C. briggsae*. BMC Biol. 10:59.

Ferrari C, et al. 2017. Ephemeral-habitat colonization and neotropical species richness of *Caenorhabditis* nematodes. BMC Ecol. 17:43.

Finn RD, et al. 2014. Pfam: the protein families database. Nucleic Acids Res. 42:D222–30.

Gonzalez de la Rosa PM, et al. 2021. A telomere-to-telomere assembly of *Oscheius tipulae* and the evolution of rhabditid nematode chromosomes. G3 Genes|Genomes|Genetics 11(1). doi: 10.1093/g3journal/jkaa020

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

Harris TW, et al. 2020. WormBase: a modern model organism information resource. Nucleic Acids Res. 48:D762–D767.

Hill RC, et al. 2006. Genetic flexibility in the convergent evolution of hermaphroditism in *Caenorhabditis* nematodes. Dev Cell. 10: 531–538.

Hillier LW, et al. 2007. Comparison of *C. elegans* and *C. briggsae* genome sequences reveals extensive conservation of chromosome organization and synteny. PLoS Biol. 5:e167.

Hoff KJ, Lange S, Lomsadze A, Borodovsky M, Stanke M. 2016. BRAKER1: unsupervised RNA-Seq-based genome annotation with GeneMark-ET and AUGUSTUS. Bioinformatics 32:767–769.

Hoff KJ, Lomsadze A, Borodovsky M, Stanke M. 2019. Whole-genome annotation with BRAKER. In: Kollmar M, editors. Gene prediction: methods and protocols. New York (NY): Springer. p. 65–95.

Hubley R, et al. 2016. The Dfam database of repetitive DNA families. Nucleic Acids Res. 44:D81–D89.

Huerta-Cepas J, Serra F, Bork P. 2016. ETE 3: reconstruction, analysis, and visualization of phylogenomic data. Mol Biol Evol. 33: 1635–1638.

Kanzaki N, et al. 2018. Biology and genome of a newly discovered sibling species of *Caenorhabditis elegans*. Nat Commun. 9:3216.

Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 30:772–780.

Kim C, et al. 2019. Long-read sequencing reveals intra-species tolerance of substantial structural variations and new subtelomere formation in *C. elegans*. Genome Res. 29:1023–1035.

Kiontke K, et al. 2004. *Caenorhabditis* phylogeny predicts convergence of hermaphroditism and extensive intron loss. Proc Natl Acad Sci U S A. 101:9003–9008.

Kiontke KC, et al. 2011. A phylogeny and molecular barcodes for *Caenorhabditis*, with numerous new species from rotting fruits. BMC Evol Biol. 11:339.

Kolmogorov M, Yuan J, Lin Y, Pevzner PA. 2018. Assembly of long error-prone reads using repeat graphs. Nat Biotechnol 37: 540–546. bioRxiv 247148.

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

Kovaka S, et al. 2019. Transcriptome assembly from long-read RNA-seq alignments with StringTie2. Genome Biol. 20:278.

Lande R. 1979. Effective deme sizes during long-term evolution estimated from rates of chromosomal rearrangement. Evolution 33: 234–251.

Lee D, et al. 2021. Balancing selection maintains hyper-divergent haplotypes in *Caenorhabditis elegans*. Nat Ecol Evol. 5:794–807.

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

Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics 34:3094–3100.

Llorens C, et al. 2011. The Gypsy Database (GyDB) of mobile genetic elements: release 2.0. Nucleic Acids Res. 39:D70–D74.

Mathé C, Sagot M-F, Schiex T, Rouzé P. 2002. Current methods of gene prediction, their strengths and weaknesses. Nucleic Acids Res. 30:4103–4117.

Muggeo VMR. 2003. Estimating regression models with unknown break-points. Stat Med. 22:3055–3071.

Nayak S, Goree J, Schedl T. 2005. fog-2 and the evolution of self-fertile hermaphroditism in *Caenorhabditis*. PLoS Biol. 3:e6.

Nigon V, Dougherty EC. 1949. Reproductive patterns and attempts at reciprocal crossing of Rhabditis elegans Maupas, 1900, and Rhabditis briggsae Dougherty and Nigon, 1949 (Nematoda: Rhabditidae). J Exp Zool. 112:485–503.

Noble LM, et al. 2021. Selfing is the safest sex for *Caenorhabditis tropicalis*. Elife 10:e62587.

Nurk S, et al. 2020. HiCanu: accurate assembly of segmental duplications, satellites, and allelic variants from high-fidelity long reads. Genome Res. 30:1291–1305.

Pertea G, Pertea M. 2020. GFF utilities: GffRead and GffCompare. F1000Res. 9. doi:10.12688/f1000research.23297.2

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26:841–842.

Ren X, et al. 2018. Genomic basis of recombination suppression in the hybrid between *Caenorhabditis briggsae* and *C. nigoni*. Nucleic Acids Res. 46:1295–1307.

Rhie A, et al. 2021. Towards complete and error-free genome assemblies of all vertebrate species. Nature 592:737–746.

Rhie A, Walenz BP, Koren S, Phillippy AM. 2020. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. Genome Biol. 21:245.

Riddle DL, Blumenthal T, Meyer BJ, Priess JR. 2011. C. elegans II. Cold Spring Harbor (NY): Cold Spring Harbor Laboratory Press.

Rockman MV, Kruglyak L. 2008. Breeding designs for recombinant inbred advanced intercross lines. Genetics 179:1069–1078.

Rockman MV, Kruglyak L. 2009. Recombinational landscape and population genomics of *Caenorhabditis elegans*. PLoS Genet. 5: e1000419.

Rödelsperger C. 2021. The community-curated *Pristionchus pacificus* genome facilitates automated gene annotation improvement in related nematodes. BMC Genom. 22:216.

Rognes T, Flouri T, Nichols B, Quince C, Mahé F. 2016. VSEARCH: a versatile open source tool for metagenomics. PeerJ 4:e2584.

Ross JA, et al. 2011. *Caenorhabditis briggsae* recombinant inbred line genotypes reveal inter-strain incompatibility and the evolution of recombination. PLoS Genet. 7:e1002174.

Ruan J, Li H. 2019. Fast and accurate long-read assembly with wtdbg2. Nat Methods 17:155–158. bioRxiv 530972.

Seidel HS, Rockman MV, Kruglyak L. 2008. Widespread genetic incompatibility in *C. elegans* maintained by balancing selection. Science 319:589–594.

Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 31: 3210–3212.

Smit AFA, Hubley R, Green P. 2015. RepeatMasker Open-4.0. 2013–2015.

Stein LD, et al. 2003. The genome sequence of *Caenorhabditis briggsae*: a platform for comparative genomics. PLoS Biol. 1:e45.

Steinbiss S, Willhoeft U, Gremme G, Kurtz S. 2009. Fine-grained annotation and classification of de novo predicted LTR retrotransposons. Nucleic Acids Res. 37:7002–7013.

Stevens L, et al. 2019. Comparative genomics of 10 new *Caenorhabditis* species. Evol Lett. 3:217–236.

Stevens L, et al. 2020. The genome of *Caenorhabditis bovis*. Curr Biol. 30:1023–1031.

Teterina AA, Willis JH, Phillips PC. 2020. Chromosome-level assembly of the *Caenorhabditis remanei* genome reveals conserved patterns of nematode genome organization. Genetics 214:769–780.

Thomas CG, et al. 2015. Full-genome evolutionary histories of selfing, splitting, and selection in *Caenorhabditis*. Genome Res. 25: 667–678.

Thompson OA, et al. 2015. Remarkably divergent regions punctuate the genome assembly of the *Caenorhabditis elegans* Hawaiian strain CB4856. Genetics 200:975–989.

Tyson JR, et al. 2018. MinION-based long-read sequencing and assembly extends the *Caenorhabditis elegans* reference genome. Genome Res. 28:266–274.

Vaser R, Sović I, Nagarajan N, Šikić M. 2017. Fast and accurate de novo genome assembly from long uncorrected reads. Genome Res. 27: 737–746.

Walker BJ, et al. 2014. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS ONE 9:e112963.

Wenger AM, et al. 2019. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. Nat Biotechnol. 37:1155–1162.

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

Yoshimura J, et al. 2019. Recompleting the *Caenorhabditis elegans* genome. Genome Res. 29:1009–1022.

Associate editor: Shu-Dan Yeh