CGCI, a new reference genome for *Caenorhabditis elegans*

Kazuki Ichikawa,[1] Massa J. Shoura,[2,8] Karen L. Artiles,[2] Dae-Eun Jeong,[2] Chie Owa,[1] Haruka Kobayashi,[1] Yoshihiko Suzuki,[1] Manami Kanamori,[3] Yu Toyoshima,[3] Yuichi Iino,[3] Ann E. Rougvie,[4] Lamia Wahba,[5] Andrew Z. Fire,[2,6] Erich M. Schwarz,[7] and Shinichi Morishita[1]

[1]*Department of Computational Biology and Medical Sciences, Graduate School of Frontier Sciences, The University of Tokyo, Chiba 277-8583, Japan;* [2]*Department of Pathology, Stanford University, Stanford, California 94305, USA;* [3]*Department of Biological Sciences, Graduate School of Science, The University of Tokyo, Bunkyo-ku, Tokyo 113-0033, Japan;* [4]*Department of Genetics, Cell Biology, and Development, University of Minnesota, Minneapolis, Minnesota 55454, USA;* [5]*Laboratory of Non-Canonical Modes of Inheritance, Rockefeller University, New York, New York 10065, USA;* [6]*Department of Genetics, Stanford University, Stanford, California 94305, USA;* [7]*Department of Molecular Biology and Genetics, Cornell University, Ithaca, New York 14853, USA*

The original 100.3 Mb reference genome for *Caenorhabditis elegans*, generated from the wild-type laboratory strain N2, has been crucial for analysis of *C. elegans* since 1998 and has been considered complete since 2005. Unexpectedly, this long-standing reference was shown to be incomplete in 2019 by a genome assembly from the N2-derived strain VC2010. Moreover, genetically divergent versions of N2 have arisen over decades of research and hindered reproducibility of *C. elegans* genetics and genomics. Here we provide a 106.4 Mb gap-free, telomere-to-telomere genome assembly of *C. elegans*, generated from CGCI, an isogenic derivative of the N2 strain. We use improved long-read sequencing and manual assembly of 43 recalcitrant genomic regions to overcome deficiencies of prior N2 and VC2010 assemblies and to assemble tandem repeat loci, including a 772 kb sequence for the 45S rRNA genes. Although many differences from earlier assemblies come from repeat regions, unique additions to the genome are also found. Of 19,972 protein-coding genes in the N2 assembly, 19,790 (99.1%) encode products that are unchanged in the CGCI assembly. The CGCI assembly also may encode 183 new protein-coding and 163 new ncRNA genes. CGCI thus provides both a completely defined reference genome and corresponding isogenic wild-type strain for *C. elegans*, allowing unique opportunities for model and systems biology.

[Supplemental material is available for this article.]

The nematode *Caenorhabditis elegans* is a model for biology ranging from mechanistic functions of individual proteins and RNAs to multicellular interactions of development and neurobiology (Brenner 1974; Corsi et al. 2015). For any organism, a key element of modern biological understanding is to sequence and characterize its genome. *C. elegans* was the first animal to have its genome sequenced in 1998 (The *C. elegans* Sequencing Consortium 1998); by 2005, its genome was considered complete and gap-free (Hillier et al. 2005). It was thus surprising when, 14 years later, our attempt to reproduce a perfect isogenic copy of the reference *C. elegans* genome instead showed that it was neither complete nor gap-free (Yoshimura et al. 2019), as discrepancies observed by others were by then also suggesting (Hillier et al. 2005; Li et al. 2015; Tyson et al. 2018).

C. elegans' original reference genome (Table 1) was generated from the standard wild-type strain, N2 (Brenner 1974). Unfortunately, N2 was probably genetically polymorphic even when it was first frozen in 1969 (Sterken et al. 2015); it continued to accumulate observable genetic polymorphisms between laboratories into the 2000s (Gems and Riddle 2000; Vergara et al. 2009); and, at this point, no frozen stock representing its original genotype exists. This lack of a frozen isogenic *C. elegans* wild-type strain with an exactly matched genome motivated our first effort to reproduce the *C. elegans* reference genome from an isogenic derivative of N2 (Table 1), which we published in 2019 as VC2010 (Yoshimura et al. 2019). VC2010 reproduced 100.3 Mb of N2 sequences with 99.98% identity but also contained an extra 1.8 Mb of genomic sequence, along with 10 genomic regions we could not assemble fully: five regions with long complex tandem repeats; tandem arrays of 5S rRNA genes (980 nt), 45S rRNA genes (7197 nt), and positioning sequence on X (pSX1) sequences (172 nt) (Nelson and Honda 1985; Ellis et al. 1986; Johnson et al. 2006); and two telomeric regions on the right ends of CHROMOSOME_I and CHROMOSOME_III. For *C. elegans* to have a completely accurate reference genome, the N2 reference assembly needed to be replaced, but the VC2010 assembly could not fully replace it.

Since 2019, long-read sequencing has improved enough to allow de novo assembly of telomere-to-telomere gap-free sequences for complex eukaryotic genomes, including the first human haploid genome (Nurk et al. 2022). Our VC2010 genome assembly relied on short Illumina reads, long Pacific Biosciences (PacBio) RSII reads, and long Oxford Nanopore Technologies (ONT) MinION reads, with respective mean read lengths of 73 nt, 8.8 kb, and 14.2 kb and with respective error rates of 0.1%, 15%–20%, and 15%–20% (Yoshimura et al. 2019). Since then, PacBio and Nanopore have improved both the length and error rate of their

[8]Present address: Phinomics, Incorporated, San Carlos, CA 94070, USA
Corresponding authors: afire@stanford.edu, ems394@cornell.edu, moris@edu.k.u-tokyo.ac.jp
Article published online before print. Article, supplemental material, and publication date are at https://www.genome.org/cgi/doi/10.1101/gr.280274.124. Freely available online through the *Genome Research* Open Access option.

Table 1. Chromosome lengths and tandem repeat lengths of different *C. elegans* genome assemblies

	CGC1 (nt)	VC2010 (nt)	Difference of CGC1 from VC2010 (nt)	N2 (nt)	Difference of CGC1 from N2(nt)
Chromosome lengths					
I	16,508,284	15,331,301	1,176,983	15,072,434	1,435,850
II	15,800,154	15,525,148	275,006	15,279,421	520,733
III	14,624,068	14,108,536	515,532	13,783,801	840,267
IV	18,555,121	17,759,200	795,921	17,493,829	1,061,292
V	22,067,386	21,243,235	824,151	20,924,180	1,143,206
X	18,785,586	18,110,855	674,731	17,718,942	1,066,644
MtDNA	13,994	13,988	6	13,794	200
Total	106,354,593	102,092,263	4,262,330	100,286,401	6,068,192
Length of 174 > 5 kb tandem repeats in each chromosome					
I	1,448,169	300,262	1,147,907	98,083	1,350,086
II	560,095	285,098	274,997	100,630	459,465
III	904,536	410,543	496,012	164,534	740,002
IV	1,151,459	353,570	797,889	113,950	1,037,509
V	1,168,580	432,266	736,314	152,978	1,015,602
X	1,111,883	448,391	663,527	137,334	974,549
MtDNA	0	0	0	0	0
Total	6,344,722	2,230,130	4,116,646	767,509	5,577,213

long reads: PacBio Sequel II generates HiFi reads with a sequencing error rate of 0.1% and a mean read length of 15–20 kb (Wenger et al. 2019), whereas Nanopore PromethION generates ultralong reads with an error rate of 1% and 10% of their total sequence output (N10) having lengths of ≥150 kb (Shafin et al. 2020). These advances have enabled complete genome assemblies for humans (Nurk et al. 2022) and other multicellular eukaryotes (Xie et al. 2024) and motivated us to assemble a comparable *C. elegans* reference genome.

Results

CGCI, an isogenic *C. elegans* strain derived from N2

As a *C. elegans* reference strain, we selected CGC1, an isogenic derivative of the original reference strain N2 (Brenner 1974) by way of VC2010 (Flibotte et al. 2010). We originally called this isogenic derivative "PD1074" while calling our genome assembly of it "VC2010" (Yoshimura et al. 2019); however, these two names were confusingly different from each other and were both difficult to remember. We thus renamed PD1074 as CGC1. With a few exceptions noted below, CGC1 should be genetically equivalent to N2; but unlike N2, it is as close to perfectly isogenic as possible and has been mass-frozen in a multitude of vials, which should make its genetics and genomics fully reproducible by *C. elegans* laboratories over many years. CGC1 is available at the Caenorhabditis Genetics Center stock center (https://cgc.umn.edu/strain/CGC1).

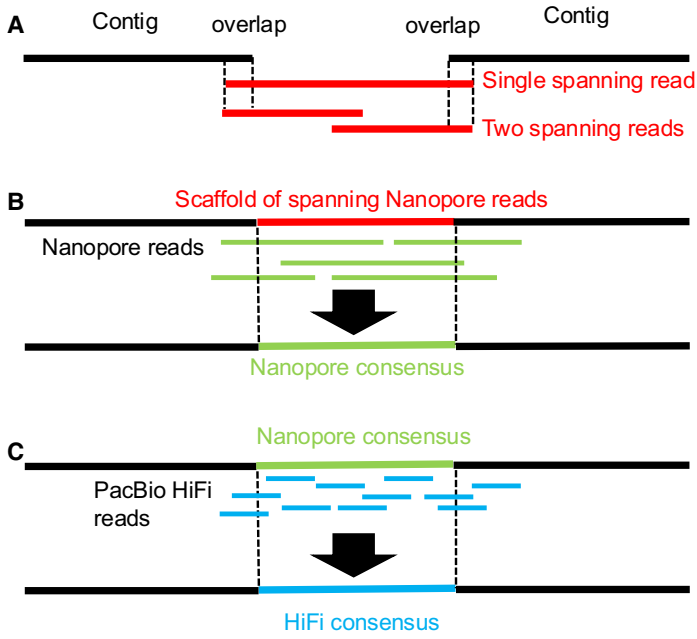
Genome sequencing and assembly

We harvested genomic DNA (gDNA) from CGC1 and sequenced it in two sets apiece: HiFi reads (Wenger et al. 2019) with 85× and 218× genome coverage, and Nanopore chemistry R9 reads (Shafin et al. 2020) with 123× and 164× coverage (Supplemental Table S1A). HiFi and Nanopore had complementary advantages: HiFi reads had a 10-fold lower error rate (0.1%) compared with Nanopore (4.2% of chemistry R9 and 1.5% of R10) (Supplemental Table S1C), whereas Nanopore reads were 7.6-fold longer (N10 of 151,427 nt in set 1) than those of HiFi (N10 of 19,847 nt in set 1) (Supplemental Table S1B). We used the first HiFi and Nanopore read sets to assemble CGC1 and used the second read sets to confirm the assembly's accuracy.

We first assembled HiFi reads into 80 contigs (Supplemental Table S2A) with HiCanu (Nurk et al. 2020) and purge_dups (Guan et al. 2020) and further reduced them to 61 nonredundant contigs by manually inspecting their order after aligning them to the VC2010 assembly (Supplemental Fig. S1). This left us with 54 gaps between neighboring contigs, all of which could be spanned by Nanopore ultralong reads. Eleven gaps could be filled with the consensus of Nanopore ultralong reads. The remaining 43 gaps were filled through manual assembly, by first using Nanopore ultralong reads to close gaps throughout the genome assembly and then using either Nanopore or HiFi reads to correct errors and generate the complete genome (Fig. 1). Figure 2 shows two difficult examples in which two Nanopore reads span neighboring tandem repeats in CHROMOSOME_I (4,715,580–4,817,023) (Fig. 2A,C) and CHROMOSOME_II (14,800,218–14,842,635) (Fig. 2B,D). All but one of the 43 gaps could be spanned by a single Nanopore read; one remaining gap could be covered by combining two neighboring Nanopore ultralong reads (Supplemental Fig. S2). We thus could find a series of alternating contigs and Nanopore ultralong reads that fully spanned each chromosome.

Error correction of closed genomic gaps

Having closed 43 gaps with 1–2 Nanopore ultralong reads in Nanopore data set 1 (Supplemental Table S1) as shown in Figure 1A, we had to correct errors of these reads by aligning all



D

Summary Statistics		
Strain	N2	CGC1
Genome Size (b)	100,286,401	106,354,593
Gap bases (b)	0	0
# contigs	7	7
Contig NG50 (b)	56,136,951	59,408,093
Gene annotation		
# Coding genes	19,972	21,238
RepeatMasker		
Repeat bases (b)	19,526,554	22,423,921
LINE (b)	475,978	480,697
SINE (b)	116,990	117,230
LTR (b)	615,612	619,954
Satellite (b)	799,243	2,296,994
DNA (b)	10,502,233	10,608,497
Simple repeat (b)	1,006,969	1,173,942
Low complexity (b)	276,660	284,140
Retroposon (b)	615,612	619,954
rRNA (b)	7,367	522,891

Figure 1. The overlap-layout-consensus approach of assembling tandem repeat regions. (*A*) The overlap-and-layout step finds one or more Nanopore ultralong reads that span a focal gap and lays out multiple reads properly. (*B*) The consensus step corrects errors in the scaffold of Nanopore reads (red), aligns other Nanopore reads (green) to the scaffold, and calculates the consensus of the aligned reads. (*C*) To further eliminate errors in the consensus sequence (green), HiFi reads (blue) are aligned to the consensus and used to generate the consensus of mapped HiFi reads. (*D*) Comparison between the N2 and CGC1 assemblies. (#) The number of contigs, coding genes, and so on. The N2 genome assembly version is WBPS19 WBcel235 (GCA_000002985.3) generated in December 2012.

Nanopore reads in data sets 1 and 2 (Supplemental Table S1) to each gap and computing the consensus of Nanopore reads (Fig. 1B). We then aligned HiFi reads to each Nanopore consensus and examined if sequencing errors could be corrected. For 25 out of 43 gaps, HiFi reads were useful in correcting sequencing errors, whereas for 13 of 43 gaps, Nanopore consensus sequences were sufficient. The remaining five of 43 gaps had complex regions with two different tandem repeats, making the process more complicated. We observed that 1.77%–3.59% of nucleotides in these five regions were corrected by the Nanopore consensus step (Table 2). However, we could further improve gaps by aligning HiFi reads to their Nanopore consensus and generating the consensus of aligned HiFi reads (Fig. 1C; Supplemental Fig. S3; Methods for details). Overall, sequencing errors in 43 gaps were corrected by using consensus of Nanopore reads, consensus of HiFi reads, or a combination of both: the column of gaps in Supplemental Table S3 shows how each gap was filled, and Supplemental Figure S4 details the read coverage distribution of each gap by Nanopore and HiFi reads.

Comparison of CGC1 to automated GALA and hifiasm assemblies

Because alternative genome assembly methods are continually being developed, we tested our CGC1 assembly against two alternative approaches. Awad and Gan (2023) have recently described GALA, a program for gap-free chromosomal long-read de novo genome assembly and have used it to produce an alternative genome assembly for *C. elegans* (Awad and Gan 2023). We compared our CGC1 assembly with their GALA assembly on four tandem repeat regions (Supplemental Fig. S5). Three regions in the GALA assembly were much shorter than the corresponding regions in the CGC1 assembly, and another region in the GALA assembly was highly inconsistent with a Nanopore ultralong read and the CGC1 assembly. This problem with the GALA assembly is presumably owing to their use of our limited PacBio and Nanopore sequence data reported in 2019 (Yoshimura et al. 2019), which was of lower quality than the new data in this study, demonstrating that the significant improvement in sequence quality of PacBio and Nanopore sequencing contributes to the high quality of CGC1 assemblies.

Cheng et al. (2021) have devised hifiasm, an assembler that can combine both HiFi and Nanopore reads into a single genome sequence. We used hifiasm 0.19.7 to assemble our two HiFi data sets with ≥20 kb reads and two Nanopore data sets with ≥50 kb reads that used chemistry R9 (Supplemental Table S1). This assembly failed to assemble complex tandem repeat regions, divided the 45S rDNA array into several smaller fragments (Supplemental Fig. S6), and output a single contig for pSX1 that was much shorter than the pSX1 region of length ~153 kb in the CGC1 assembly (Supplemental Table S2B; Supplemental Fig. S7). This demonstrates how highly tandemly repeated regions are still difficult to assemble automatically and require manual inspection.

Tandem repeats newly found in the CGC1 assembly

The CGC1 genome assembly is longer than the earlier VC2010 assembly (Yoshimura et al. 2019) by 3.8 Mb; 96% of this difference consists of 174 tandem repeats ≥5 kb in size (Fig. 1D; Table 1; Supplemental Table S3; Supplemental Data Files 1–3). Tandem repeats in CGC1 are noticeably larger than those also found in the VC2010 assembly; tandem repeats of ≥50 kb in size are observed only in the CGC1 assembly (Supplemental Fig. S8A–C), presumably because previous methods underestimated them.

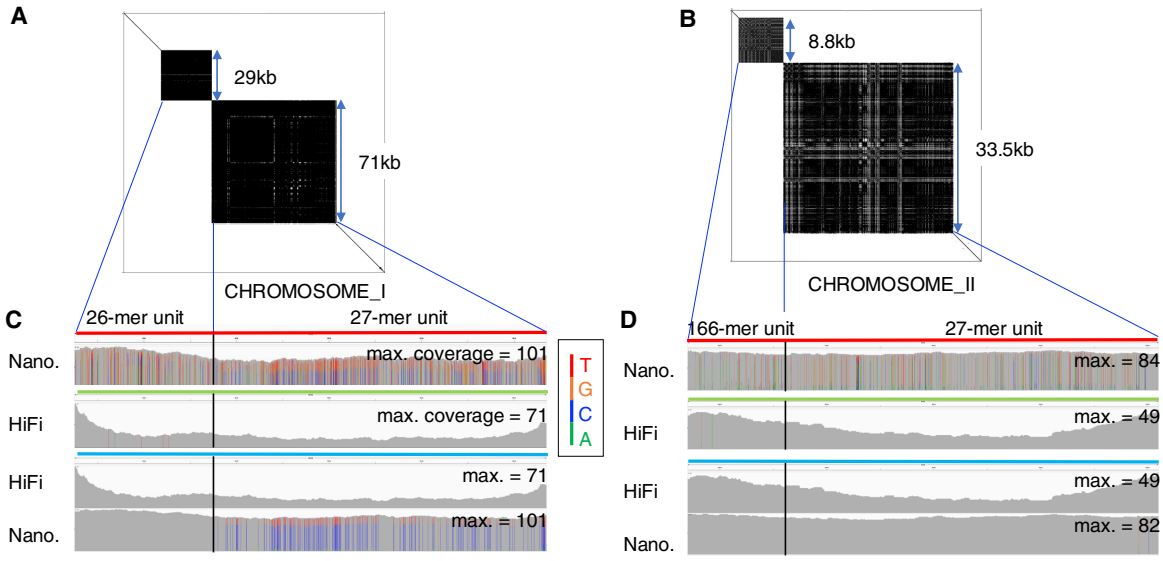


Figure 2. Polishing two complex tandem repeat regions in CHROMOSOME_I and CHROMOSOME_II. (*A*,*B*) Self-to-self dot plots of two Nanopore reads that span neighboring tandem repeats in CHROMOSOME_I (4,715,580–4,817,023; *A*) and CHROMOSOME_II (14,800,218–14,842,635; *B*). (*C*) The *top* row shows a spanning Nanopore read (colored red) and the coverage of other Nanopore reads mapped to the spanning read in CHROMOSOME_I. The bar for each base represents the distribution of aligned bases in the consensus. If one nucleotide accounts for >80%, the bar is colored gray; otherwise, the four nucleotides A, C, G, and T are green, blue, orange, and red, respectively. The second row shows coverage of HiFi reads mapped to the Nanopore consensus (light green). The third and fourth rows present coverages of HiFi reads and Nanopore reads to the HiFi consensus (light blue). Bars in the third row are gray, showing the consistency of HiFi reads with the HiFi consensus. The fourth row shows that there are many bars with C's (blue) as the majority and T's (red) as the minority, but the base in the consensus of all bars is C. (*D*) Read coverage of a spanning Nanopore read, Nanopore consensus, and HiFi consensus for complex tandem repeat regions in CHROMOSOME_II.

Of these 174 ≥ 5 kb tandem repeats, 50 were present in 43 gaps (seven gaps had two neighboring repeats), whereas the others are present in contigs assembled from HiFi reads by HiCanu. We tested how accurately we had assembled these repeat-containing contigs by mapping Nanopore ultralong reads, determining consensuses of the mapped Nanopore reads, and comparing the consensuses to the contigs. Nanopore read coverage was uniform in each tandem repeat and no significant structural discrepancies were found (Supplemental Fig. S4).

Of these 174 ≥ 5 kb tandem repeats, some shared identical or near-identical units in common but were located on different chromosomes. A prominent instance of this was a set of 12 regions sharing the 27-mer tandem repeat 5′-ACTCTCTGTGGCTTCC CACTATATTTT-3′, which motivated us to analyze their evolutionary proximity (Fig. 3). Despite the evolutionary proximity of the tandem repeat regions in two groups of units, they are located far apart on different chromosomes. Supplemental Figure S9 and Supplemental Table S4 show the distribution of various units inside each region, implying that detailed analysis on the distributions of unit variants is informative to understand evolution of tandem repeats.

Assembly and analysis of the pSXI and 5S rDNA tandem repeat regions

It has been challenging to assemble accurate sequences from genomic arrays of pSX1 (172 nt in size), 5S rDNA (976 nt), and 45S rDNA (7197 nt) that form extremely large repetitive loci in the *C. elegans* genome (Nelson and Honda 1985; Ellis et al. 1986; Johnson et al. 2006). In CGC1, we found that three ultralong reads and one ultralong read, respectively, spanned the pSX1 and 5S rDNA arrays of sizes 153 kb and 212 kb (Supplemental Fig. S10A, B). We corrected sequencing errors in their spanning reads by aligning Nanopore reads and PacBio HiFi reads to the two spanning reads and determining the consensus sequences of aligned reads (Supplemental Fig. S10C,D). In the resulting error-corrected pSX1 and 5S rDNA arrays, we identified 887 and 198 copies,

Table 2. Error correction by using the Nanopore and PacBio consensuses

Chromosome	Starting nucleotide	Ending nucleotide	Total length of tandem repeats	Nanopore corrections	HiFi corrections
I	4,715,580	4,817,023	101,443	2627 (2.59%)	24 (0.024%)
II	14,800,218	14,842,635	42,417	1093 (2.58%)	6 (0.014%)
III	7,856,863	8,031,900	175,037	5439 (3.11%)	25 (0.014%)
X	4,508,408	4,592,148	83,740	3007 (3.59%)	NA
X	5,581,121	5,673,280	92,159	1633 (1.77%)	NA

Nucleotide coordinates of these five gap regions are given for the final CGC1 genome assembly Corrections for each read type (Nanopore or HiFi) include the total number of corrected bases, insertions, and deletions, and their ratio to the total length (in parentheses). (NA) No HiFi corrections are made in CHROMOSOME_X.

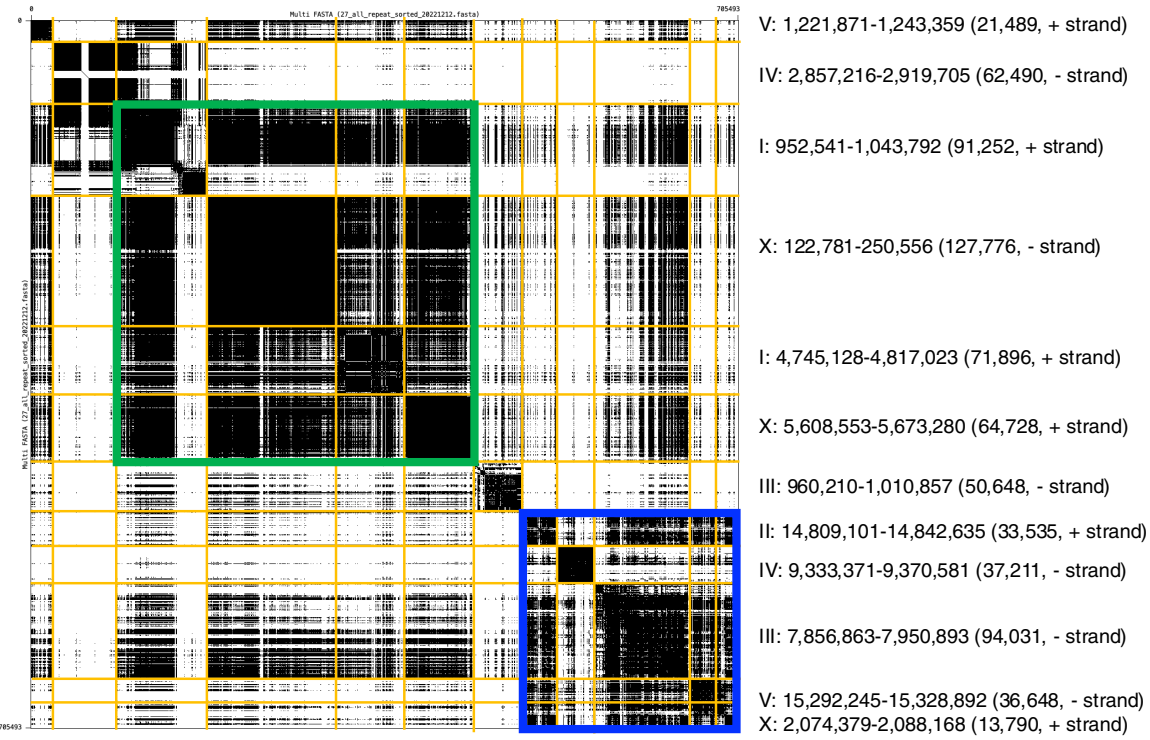


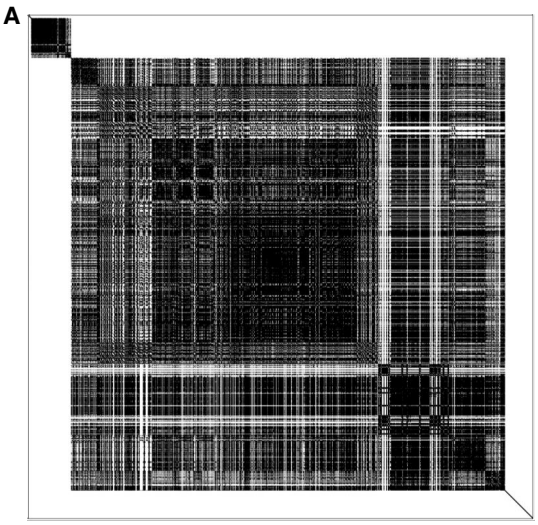
Figure 3. Similarity of prominent 27-mer tandem repeat regions. Self-to-self dot plot of 12 prominent tandem repeat genomic regions sharing a common 27-mer repeat unit. For example, the *top* row shows a 21,489 nt genomic region of CHROMOSOME_V (1,221,871–1,243,359) on the + (Watson) strand. Dots represent perfect matches of length 54 nt. For example, the dot plot for the fourth genomic region of CHROMOSOME_X (122,781–250,556) is very dense and appears black owing to the high similarity between the 54-mers. In contrast, white areas indicate discordance. Regions are grouped by sequence similarity rather than their genomic position. The green box contains four genomic regions that share the most frequent version of the 27-mer unit (5′-ACTCTCTGTGGCTTCCCACTATATTTT-3′), which we call type1. The blue box contains five genomic regions that share many copies of both type1 and another version of the 27-mer, type2 (5′-ACTCTCTGTGGCTTCCCACCATATTTT-3′), that has one (underlined) T-to-C substitution with respect to type1.

respectively, of pSX1 and 5S rDNA elements (Supplemental Table S5). In the 5S rDNA array, 211 copies of SL1 gene were also found. In the VC2010 assembly (Yoshimura et al. 2019), these had been underestimated to have 167 and 144–145 copies, respectively. This discrepancy might be owing to our lack of ultralong reads when assembling VC2010. Alternatively, it might reflect real genetic change: Tandem repeat lengths at the rDNA locus of *C. elegans* can vary dynamically over time (Wahba et al. 2021), and such variation might at least partially account for differences between our earlier VC2010 and later CGC1 genome assemblies. An independent assembly of the 5S rDNA locus with Nanopore reads from the N2 strain by Ding et al. (2022) estimated 167 consecutive 5S rDNA units. These results demonstrate the importance of Nanopore ultralong reads for properly assembling large tandemly repeated genomic sequences.

The complete sequences of the pSX1 and 5S rDNA arrays provide an opportunity to study how they evolved (Fig. 4A,B; Supplemental Fig. S11). In both arrays, some variants of the repeated element appear to preferentially occur near one another significantly (according to the Wald–Wolfowitz runs test), suggesting that they might have expanded through replication slippage of a single progenitor variant. The 153 kb pSX1 array in CHROMOSOME_X was homologous to another 115 kb pSX1 array in CHROMOSOME_V (Supplemental Fig. S12). In the CHROMOSOME_X array, the most frequent 11 pSX1 variants matched the reference pSX1 element by >97%, and each variant had a unique set of substitutions and insertions (Fig. 4B). In contrast, in the CHROMOSOME_V array, the most frequent 10 variants match the reference by ≤92%, and they lacked unique substitutions or deletions that would allow individual variants to be unambiguously distinguished (Supplemental Fig. S13). One possible explanation for this pattern is that the CHROMOSOME_X array arose by more recent transposition and amplification of a simpler pSX1 segment from the CHROMOSOME_V array, which would give elements in the CHROMOSOME_X array less time to diverge from one another.

Assembly and analysis of the 45S rDNA tandem repeat region

To assemble the 45S rDNA array encoding 18S, 5.8S, and 28S rRNAs, we first searched the 45S rDNA representative repeat unit of 7197 nt (Supplemental Table S5) for single-nucleotide variants (SNVs) that could distinguish the positions of each 45S rDNA unit variant within the array. We aligned 3322 PacBio HiFi reads to the 45S rDNA representative repeat unit. Because the HiFi reads were long enough for the 45S rDNA unit to occur approximately twice, HiFi read coverage in the 45S rDNA unit averaged 6691 at 7197 positions. In six of the 7197 positions, minor SNVs were detected 60 or more times (Fig. 5A); we thus used these six SNVs as markers to assemble Nanopore reads to construct a 45S rDNA array. To do this reliably with Nanopore reads that had an error rate of ~5% for the six SNVs, we searched for positions that perfectly matched 11 bases centered on each of the six SNV sites. We also used a HiFi-validated insertion of GTCC at position 2811 as a



B

# Occ.	Occ. Ratio	# Mis.	Match Ratio	Substitutions	Insertions	P
26	3.48	5	97.09	171A	51TCT,172A	
23	2.58	2	98.84	171A	172A	*
78	8.75	3	98.26		51TCT	***
36	4.04	0	100.00			***
105	11.78	1	99.42	114T		**
74	8.31	2	98.84	23G, 114T		
18	2.02	2	98.84	58C, 131T		**
119	13.36	1	99.42	84T		**
41	4.60	3	98.26	90G, 91A,170T		
38	4.26	2	98.84	84T,170T		***
26	2.92	3	98.26	72G, 84T, 170T		

Figure 4. Structure of the pSX1 array. (*A*) Self-to-self dot plot of a region with the 153 kb pSX1 array in CHROMOSOME_X (441,085–594,508). Dots represent perfect matches of substrings of length 100 nt. (*B*) On the *left* side of the table, the phylogenetic tree shows the proximity of variants in terms of sequence similarity. On the *right* side of the table, the colored bars indicate the locations of the occurrences of each variant. Characteristics of 11 frequent variants of the reference pSX1 (172 nt; orange): the number of occurrences of each variant (# Occ.), the ratio of number of occurrences to total (Occ. ratio), number of mismatches with the reference pSX1 (# Mis.), percentage match with the reference (match ratio), positions of base substitutions (e.g., 171A means the base at position 171 is substituted with *A*), positions with insertion bases (e.g., 51TCT shows TCT is inserted at position 51), and the statistical significance codes for *P*-values. (***) $P \leq 0.1\%$, (**) $P \leq 1\%$, and (*) $P \leq 5\%$ such that each variant occurs adjacent to each other preferentially according to the Wald–Wolfowitz runs test.

marker. We thus extracted Nanopore reads with the six SNVs from both Nanopore ultralong read data sets (Supplemental Table S1), aligned them so that the distances of adjacent SNVs and insertions within reads were consistent between aligned reads (Fig. 5B), and calculated an initial consensus from these aligned Nanopore reads (Fig. 5C). To correct persistent errors, we aligned ≥100 kb Nanopore reads in the two Nanopore read data sets to the initial consensus and recalculated the final consensus of the aligned reads.

This yielded a final assembly of the 45S rDNA array with 107 units totaling 772 kb. This resembles both an independent size estimate of 799 kb from our own PacBio HiFi reads (Methods) and another study that used Nanopore sequencing to estimate a size of 705–820 kb for the 45S rDNA array in gDNA from a different laboratory population derived from N2 (Ding et al. 2022). We further confirmed the accuracy of the variants in our assembled 45S rDNA assembly for CGC1 by mapping RNA-seq reads to the assembly (Wahba et al. 2021) and verifying that they matched the six SNVs and GTCC insertion (Supplemental Table S6). We also tested the hypothesis that all 45S rDNA repeats occurred in a single array at the right end of CHROMOSOME_I by checking 497 Nanopore reads in the Nanopore data set 1 that had BLASTN hits of 45S rDNA and were ≥30 kb in size. After removing three chimeric reads, we found that the remaining 494 reads were consistent with the hypothesis: 442 of the 494 reads were filled with 45S rDNA occurrences; 36 had telomeric repeats $(GGCTTA)_n$ at one end; and 16 had the unique region before the 45S rDNA array at one end (Supplemental Table S7).

Finally, because changes in the copy number of ribosomal RNA genes are observed in other species such as budding yeast and humans (Hori et al. 2023), we investigated how the 45S rDNA array of CGC1 differs from arrays in other *C. elegans* strains. For this, we analyzed publicly available PacBio HiFi read data sets from two strains "DLW N2 Bristol" and "DLW CB4856 [Hawaiian strain]" (Bush et al. 2025), and from two strains named ALT1 and ALT2 derived from N2 and CB4856 (Lee et al. 2023), according to the same procedure of analyzing CGC1. Because Nanopore ultralong reads were not available for these four strains, we could not completely assemble their 45S rDNA arrays. However, we reliably detected SNVs within the 7197 nt 45S rDNA representative repeat unit; three of the six SNVs found in CGC1 were also found in the DLW N2 population, whereas no SNVs were present in the other three strains (Supplemental Table S8). On the other hand, DLW

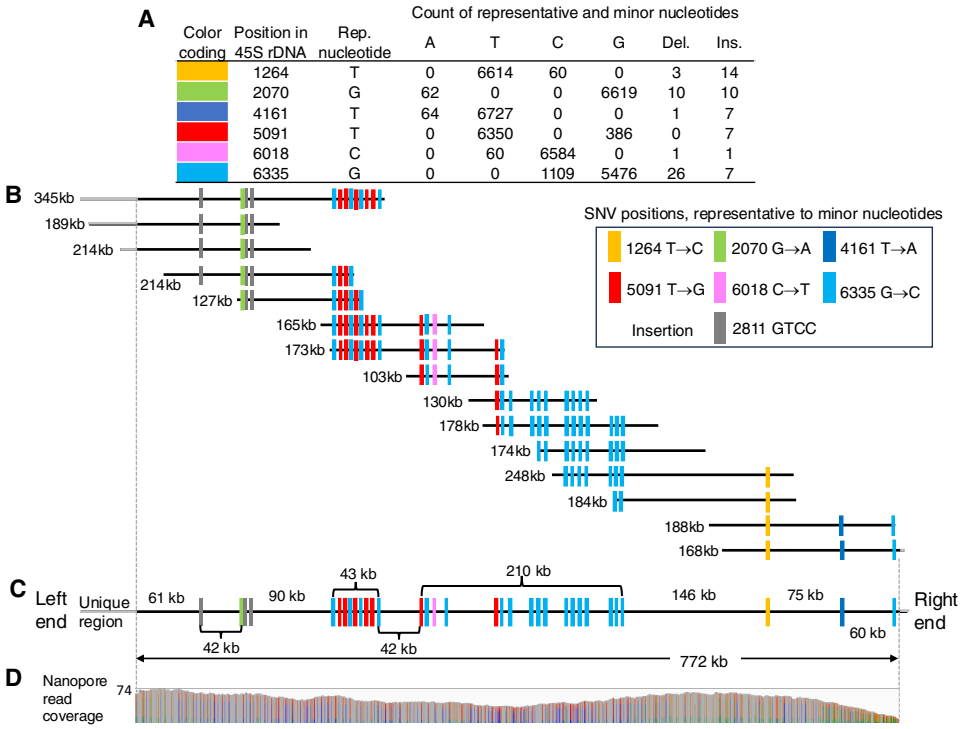


Figure 5. 45S rDNA array at the right end of CHROMOSOME_I. (*A*) Single-nucleotide variants (SNVs) detected in the 45S rDNA representative repeat unit of size 7197 nt; 3322 PacBio HiFi reads were aligned to the 45S rDNA representative repeat unit. Because each HiFi read was long enough for the 45S rDNA unit to occur approximately twice, HiFi read coverage in the 45S rDNA unit averaged 6691 at 7197 positions. The table shows SNVs whose minor nucleotides are detected 60 or more times in PacBio HiFi reads. (*B*) Nanopore reads are aligned so that distances of adjacent SNVs and insertions within Nanopore reads are consistent between aligned reads. (*C*) Consensus of aligned Nanopore reads, an assembly of the 45S rDNA array with 107 units. Each position in the consensus is covered by two or more Nanopore reads shown in *B*. (*D*) Histogram displays read coverage for each position within the consensus by Nanopore reads >100 kb in length.

CB4856, ALT1, and ALT2 each had SNVs absent from in the CGC1 and DLW N2 strains. Thus, SNVs in the 45S rDNA array evolutionarily diverged between wild-type strains of *C. elegans*.

Telomeric repeats on the left end and right end of each chromosome

Because CGC1 included complete assemblies of the rDNA array and a second subtelomeric repeat region, it was also possible (unlike in the earlier VC2010 assembly) to fully assemble telomeric repeats at both ends of all chromosomes of CGC1. To illustrate this, Supplemental Figure S14 displays dot plots between pairs of 10 kb genomic sequences taken from the left end and right end of each chromosome in CGC1, and Supplemental Table S9 shows the starting and ending positions of each telomeric repeat.

The CGC1 genome remains dynamic

Although frozen stocks of CGC1 can be assumed to have a constant genome composition, it remains important to recognize that natural mutagenic processes will yield both individual-to-individual and temporal genetic variation in any population. Exemplifying this, an examination of sequencing data from different populations of strain CGC1 used in this work revealed a consistent site of genetic variation at position 9613 of the mitochondrial genome. This site is heteroplasmic, with a major allele (9613G) representing the majority of reads in the initial data sets and a minor allele (9613T) shown in a small fraction of sequencing reads (0%–4.56%) (Supplemental Table S10). This fraction increased in samples derived more recently from animals that were propagated in the laboratory through multiple generations. Although we do not know whether this allele varied because of adaptive selection or genetic drift, these results exemplify the biological dynamism unavoidable by any genome reference and by CGC1 in particular.

Genetic contents of CGC1

Protein-coding and ncRNA genes of *C. elegans*, along with repetitive DNA elements and pseudogenes, have been identified by manual curation of its N2 reference genome over the past 25 years (Sternberg et al. 2024). For the new CGC1 reference genome to be useful to biologists, it must be annotated in a way that preserves as much of the older N2 annotation content as possible while also indicating new genes missing in the earlier N2 assembly. To do this, we used Liftoff (Shumate and Salzberg 2021) to map N2 genes onto the CGC1 assembly (i.e., to lift the genes over from the N2 to the CGC1 assemblies; Supplemental Table S11; Supplemental Data Files 4, 5) and determined which genes kept their contents unchanged in CGC1. We also used AUGUSTUS (Stanke et al. 2008) and StringTie2 (Kovaka et al. 2019) to predict genes and transcripts in CGC1 (Supplemental Tables S12, S13; Supplemental Data Files 6–9) and identified which predictions were most likely to be genuinely new, with StringTie2 transcripts providing evidence for the reality of AUGUSTUS predictions. To provide further evidence for predictions, we mapped mass spectrometric data from three surveys of the *C. elegans* proteome (Xia et al. 2018; Müller et al. 2020; Ceron-

Noriega et al. 2023) to N2 and AUGUSTUS protein-coding gene sets. To define CGC1 genomic coordinates for sequences that either are shared with N2 (Supplemental Data File 10) or are unique to CGC1 (Supplemental Data File 11) to which genes or transcripts could be assigned, we aligned the two assemblies and identified uniquely aligned blocks between them with MUMmer4 (Marçais et al. 2018). Table 3 details our results for N2 genes mapped onto the CGC1 genome, protein-coding genes predicted with AUGUSTUS, and transcripts predicted with StringTie.

Of 19,972 nuclear protein-coding genes in N2 (WS292), 19,790 (99.1%) of N2 protein-coding genes encoded at least one unchanged product after mapping to CGC1. There were 46 protein-coding genes of N2 that, when lifted over to CGC1, overlapped AUGUSTUS predictions that also encoded exons in local blocks of CGC1-specific gDNA; these may represent additional exons of the mapped N2 genes that would be invisible without the CGC1 genome (Supplemental Table S11). Another 107 N2 protein-coding genes showed lost or altered translation when mapped to the CGC1 assembly; of these, 101 (94.4%) overlapped AUGUSTUS predictions that should enable their structures to be corrected in CGC1 by WormBase curators. Mapping of ncRNA genes and pseudogenes to CGC1 was also successful: Of 24,788 ncRNA genes in N2, 24,591 (99.2%) encoded at least one transcript identical to N2; for 2131 N2 pseudogenes, 2092 (98.2%) had at least one transcript identical to N2 (Supplemental Data File 12). Note that some genetic differences between the N2 and CGC1 genomes may be real, having arisen during the derivation of the CGC1 strain from the N2 strain; for instance, CGC1 carries a spontaneous partial deletion of *alh-2* (Maydan et al. 2010), and *alh-2* fell into our set of 83 genes with altered translation products. In addition, some N2 pseudogenes might correspond to functional genes in CGC1, because of nucleotide differences in CGC1 that restore the pseudogene to function (Stewart et al. 2005).

Long-read sequencing allows the assembly of tandemly repeated genomic regions that would otherwise be omitted by older technologies (Alkan et al. 2011; Denton et al. 2014; Nurk et al. 2022), and this can reveal previously unobserved tandemly repeated genes. We found that 34 protein-coding genes in N2 had tandemly repeated copies in CGC1, totaling 60 extra genes (Supplemental Table S11). Notably, the telomerase-associated gene *pot-3* (Yu et al. 2023) had seven tandem copies in CGC1 (Supplemental Table S11). Similarly, we found that 117 N2 ncRNA genes were tandemly repeated in CGC1 with 630 extra genes; most of these were copies of genes encoding 5S and 5.8S ribosomal RNAs (such as *rrn-2*, *rrn-4*, and *sls-1*).

Of 21,238 protein-coding genes predicted de novo with AUGUSTUS in the CGC1 assembly (Table 3), 19,232 (90.6%) overlapped 19,427 N2 protein-coding genes; thus, AUGUSTUS repredicted 97.6% of N2 protein-coding genes lifted over into the CGC1 assembly. Independently, we predicted 23,249 genomic loci encoding StringTie2 transcripts assembled from ribodepleted RNA-seq data (Wang et al. 2022). These transcripts overlapped with (and thus confirmed) known or predicted N2 or AUGUSTUS protein-coding genes and with known N2 ncRNA genes (Table 3), including long ncRNA genes such as *lep-5* (Kiontke et al. 2019), *linc-4* (Wei et al. 2019; Ishtayeh et al. 2021), and *tts-1* (Essers et al. 2015). Of the remaining 3868 StringTie2 loci not overlapping these gene sets, 163 at least partly overlapped CGC1-specific gDNA and did not encode any ncRNA motifs from Rfam (Kalvari et al. 2021); these may represent novel ncRNA genes of CGC1 (Supplemental Table S14).

Of the 2006 AUGUSTUS predictions not overlapping N2 genes, 1779 were confined to N2-shared regions of gDNA, but 227 at least partially overlapped CGC1-specific gDNA. From these 227, we used several criteria (Methods) to identify 183 possible new genes with varying degrees of credibility (Supplemental Table S15): 150 had evidence for transcription (by overlapping StringTie2 transcripts), and 13 also had evidence for translation (by having protein spectra). Of the 33 non-StringTie2-associated genes, five showed orthology to genes in N2, *Caenorhabditis nigoni*, or *Caenorhabditis remanei*. From the 183-gene set, 77 genes encoded a variety of proteins including a G protein-coupled receptor, an MFP2 motility protein, an MPV17 peroxisomal protein, a serine–threonine protein kinase, and a SIR2 sirtuin. The remaining 106 genes were tandem copies encoding a single putative ART2/RRT15 domain-containing protein with a coding sequence embedded in, but antisense to, tandemly repeated 26S ribosomal RNA genes; they overlapped with StringTie2 transcripts that were assembled from ribodepleted RNA-seq data, and they partially overlapped an ncRNA antisense to 26S rRNA (F31C3.14) identified by the modENCODE consortium (Lu et al. 2011). Homologs of these putative *C. elegans* ART2/RRT15 genes have been reported in diverse eukaryotes, including one *Saccharomyces cerevisiae* homolog (RRT15) that is also encoded antisense to rRNA and whose mutation lowers RNA polymerase I transcription (Hontz et al. 2009). At least some of this 183-gene set are likely to be real, but defining their true status will require further analysis.

As a byproduct of predicting novel protein-coding genes and ncRNA transcripts in CGC1-specific gDNA, we also predicted genes and transcripts in CGC1 sequences shared with N2. WormBase curators have spent decades evaluating gene predictions in N2 and found many to be repetitive elements or pseudogenes. Thus, we treated these new predictions in N2-shared genomic sequence with caution and disqualified them for overlaps with N2 pseudogenes or repetitive DNA elements. From 1779 AUGUSTUS genes in N2-shared genomic sequences, this left 314 possibly genuine ones (Supplemental Table S16). These 314 genes primarily encoded small proteins (median, 94 residues; range, 66–1218 residues). In addition to repredicting 20.9% of known N2 ncRNA genes, StringTie also predicted 3693 transcripts (Supplemental Table S17) that fell into N2-shared genomic sequences but did not overlap known genes or pseudogenes, and had no ncRNA motifs from Rfam (Kalvari et al. 2021). Like new gene predictions in CGC1-specific gDNA, new predictions in N2-shared gDNA are possibly but not certainly real.

Transcription of long tandem repeat regions

In *C. elegans*, TRs could be transcriptionally active either because they fall inside regions of transcribed genes that harbor them or because the TRs are themselves actively transcribed as noncoding RNA (ncRNA)–encoding genes. To address the former question, we examined the distribution of TRs in introns of protein-coding genes mapped from the N2 genome (Supplemental Table S18; Supplemental Fig. S5D). Some introns were largely filled with TRs, but a number of introns were almost free from TRs. For example, among introns with a length of ≥1 kb, 113 consisted of ≥80% TR sequences, whereas 432 consisted of ≤20% (Supplemental Fig. S5E). Thus, transcription of intronic TRs is commonplace, and some large introns tolerate being heavily occupied by TRs, but most large introns are predominantly non-TR sequence. To test further whether the 174 TRs of length ≥5 kb were transcribed, we mapped onto the CGC1 genome publicly available Nanopore RNA-seq data from different tissue types (mixed-stage embryo; L1, L2, L3, L4 larva; young adult; and gravid adult) in two

Table 3. Details of N2 genes, AUGUSTUS protein-coding genes, and StringTie transcripts

Gene category	Gene number
All protein-coding genes in N2	19,972
Encoded protein spectra	9508 (47.6% of all)
Could not be mapped onto CGC1 genome	75
Could not be mapped; micropeptides	49
Could not be mapped; micropeptides with StringTie2 matches	24
Could not be mapped; had AUGUSTUS orthologs	26
Could not be mapped; had AUGUSTUS orthologs and StringTie2 matches	14
Could not be mapped; had unique AUGUSTUS orthologs	15
Could not be mapped; encoded protein spectra	3
Mapped onto CGC1 genome	19,897
Mapped; overlapped StringTie2 transcripts	18,755 (93.8% of mapped)
Mapped; overlapped AUGUSTUS genes	19,427 (97.6% of mapped)
Overlapped AUGUSTUS genes with CGC1-specific exons	46
Could not be translated after mapping	24
Not translated but overlapped AUGUSTUS genes	22 (91.7% of 24)
Could be translated but only with an altered product	83
Only altered translation but overlapped AUGUSTUS predictions	79 (95.2% of 83)
Encoded at least one protein identical to N2	19,790 (99.1% of all)
All ncRNA genes in N2	24,788
Could not be mapped onto CGC1 genome	3
Mapped onto CGC1 genome	24,785
Mapped; overlapped StringTie2 transcripts	5173 (20.9% of mapped)
Could be transcribed but only with an altered product	194
Encoded at least one transcript identical to N2	24,591 (99.2% of all)
All pseudogenes in N2	2131
Could not be mapped onto CGC1 genome	4
Mapped onto CGC1 genome	2127
Could be transcribed but only with an altered product	35
Encoded at least one transcript identical to N2	2092 (98.2% of all)
All AUGUSTUS protein-coding genes in CGC1 genome	21,238
Encoded protein spectra	9640 (45.4% of all)
Overlapped StringTie2 transcripts	19,746 (93.0% of all)
Overlapped N2 protein-coding genes	19,232 (90.6% of all)
Did not overlap N2 protein-coding genes	2006 (9.4% of all)
No N2 prot.-cod. gene overlap; overlapped only N2 DNA	1779
Met criteria for new genes; overlapped only N2 DNA	314
Met criteria; only N2 DNA; overlapped StringTie2 transcripts	201 (64.0% of 314)
Met criteria; only N2 DNA; encoded protein spectra	12 (3.8% of 314)
No N2 prot.-cod. gene overlap; overlapped CGC1-specific DNA	227
Met criteria for new genes; overlapped CGC1-specific DNA	183
New genes; CGC1-specific DNA; overlapped StringTie2 transcripts	150 (82.0% of 183)
New genes; CGC1-specific DNA; encoded protein spectra	13 (7.1% of 183)
All genomic loci encoding StringTie2 transcripts in CGC1	23,249
Overlapped N2 protein-coding genes	16,961 (73.0% of all)
Overlapped N2 ncRNA genes	2745 (11.8% of all)
Overlapped AUGUSTUS protein-coding genes	18,059 (77.7% of all)
Not overlapping N2 or AUGUSTUS genes	3868 (16.6% of all)
No N2/AUGUSTUS overlap; overlapped only N2 DNA; did not encode any ncRNA motifs from Rfam	3693
No N2/AUGUSTUS overlap; overlapped CGC1-specific DNA; did not encode any ncRNA motifs from Rfam	163

We mapped nuclear (nonmitochondrial) N2 genes from the WS292 release of WormBase (Sternberg et al. 2024) to the CGC1 assembly as three sets: protein-coding, ncRNA, and pseudogenes. Of the 75 unmapped N2 protein-coding genes, 49 encoded micropeptides of nine to 15 amino acids (Olexiouk et al. 2018) that made them difficult to map. The other 26 unmapped N2 genes, which encoded proteins of 100 or more amino acids, all had orthologs among the AUGUSTUS gene predictions; of these, 15 were unique (1:1), whereas 11 were two to 13 AUGUSTUS genes each. We tested all unmapped genes for 100% full-length matches to StringTie2 transcripts in the CGC1 genome.

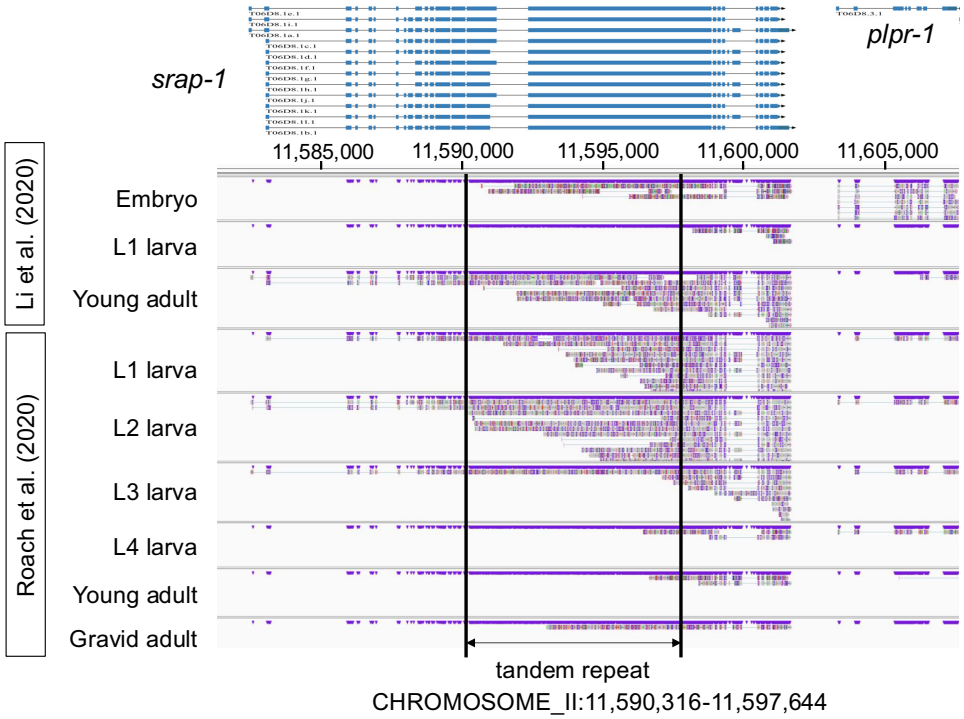


Figure 6. Transcription from a tandem repeat in various tissue types. This shows an alignment of long RNA-seq reads to a tandem repeat in CHROMOSOME_II (nt 11,590, 316–11,597, 644). The *topmost* rows show mRNA isoform splicing for genes in the region, with *srap-1* containing the tandem repeat. (*Below*) Three tracks show alignments of tissue type data from Li et al. (2020) (NCBI Gene Expression Omnibus [GEO; https://www.ncbi.nlm.nih.gov/geo/], accession number GSE130044); six additional tracks show tissue type data from Roach et al. (2020) (European Nucleotide Archive [ENA; https://www.ebi.ac.uk/ena/browser/home] accession number PRJEB31791). The alignment on tandem repeats and introns in *srap-1* was consistently observed in different tissue types, even though the RNA-seq data were collected in two independent studies, supporting the presence of transcription.

independent studies (Li et al. 2020; Roach et al. 2020). In four of the 174 long tandem repeats of length ≥5 kb (Supplemental Table S3), we observed RNA-seq reads from different tissue types that mapped onto the TRs with alternative splicing patterns outside the tandem repeats (Fig. 6; Supplemental Fig. S15). Three of these instances fall into exons or introns of protein-coding genes; a fourth instance may be termination of mRNAs from adjacent genes. The possibility remains that TRs are specifically transcribed into ncRNAs (Subirana and Messeguer 2021), as is known to occur for some loci in the *Drosophila melanogaster* genome (Biswas et al. 2024), but our long RNA-seq mappings could also be consistent with TR transcription in *C. elegans* being a byproduct of mRNA transcription.

Discussion

Using high-accuracy PacBio HiFi and ultralong Oxford Nanopore long-read sequencing, as well as augmenting our initial HiCanu assembly with manual assembly of 54 tandemly repeated regions, we have generated the first telomere-to-telomere gap-free genome assembly for the model organism *C. elegans*. This assembly encompasses the existing N2 reference genome and reproduces 99% of gene structures in N2 but adds 6 Mb of sequences. In addition to TRs, these added sequences encode additional interstitial exons for 46 protein-coding genes, 13–183 entirely new protein-coding genes, up to 163 new ncRNAs, and a 772 kb 45S rRNA gene array.

We found that automated general-purpose assemblers could generate reliable large contigs for much of the CGC1 genome, but neither HiCanu nor hifiasm could correctly assemble CGC1's longest TRs. Because *C. elegans* lies toward the lower end of genomic complexity for multicellular eukaryotes (https://www.genomesize.com/), the putative TR sequences in automatic assemblies for other eukaryotic genomes should be viewed with caution; it is likely that large TRs like those of *C. elegans* exist in many other genomes but have been underassembled or missed entirely. Here, we have proposed a strategy of covering long TRs with Nanopore ultralong reads and then aligning PacBio HiFi reads to the ultralong reads to eliminate sequencing errors (Fig. 1). When completing very long tandem repeats (e.g., the 45S rDNA array), we needed first to find positional markers manually and then to overlap ultralong reads according to those positional markers to cover an entire tandem repeat (Fig. 5). Future versions of genome assembly programs may find it possible and useful to automate this process.

The CGC1 genome assembly, with its matched clonal strain, should aid future *C. elegans* research in several ways. The most pragmatic of these is accuracy. When individual laboratories use classical and molecular genetics to dissect molecular and biological functions, they rely on their reference genome being completely correct, and since 2005, the general assumption has been that the N2 reference genome is completely correct (Hillier et al. 2005). We show that this is not entirely true: 1% of N2 genes have some uncertainty that makes them impossible to map automatically into CGC1 or they have cryptic exons only present in CGC1 or they have tandem repeat copies that the N2 assembly fails to capture. The number of additional protein-coding genes encoded exclusively by CGC1 is small and uncertain, but it is not zero. Modern forward genetics of *C. elegans* often involves chemical mutagenesis and phenotypic screening followed by

whole-genome sequencing to identify candidate alleles (Doitsidou et al. 2016): Isogenicity of the CGC1 strain with completeness of the CGC1 genome should make mutant hunts more reliable and will both simplify and improve genetic and genomic studies.

CGC1 should also aid *C. elegans* research in population genomics, chromosomal organization, and the functions of highly repetitive genomic loci. Genetic polymorphism in wild isolates of *C. elegans* should be more reliably assigned to SNVs and genomic islands of elevated diversity in the CGC1 assembly than in N2 (Crombie et al. 2019; Lee et al. 2021). Even for putatively identical wild-type strains, it will be important to use CGC1 to develop a pangenome of N2 sequences in the future. Global traits of *C. elegans* chromosomes such as meiotic recombination rates and the densities of various genomic elements should be more correctly ascertained (Carlton et al. 2022). The full sequencing of 45S rDNA array reported in this study should help determine the mechanism of maintaining rDNA clusters in *C. elegans*, as previously done in *S. cerevisiae* (Kobayashi and Ganley 2005); it might also clarify whether cellular senescence is correlated with rDNA stability, as suggested in mammalian cells (Hori et al. 2023). For many other TRs in *C. elegans*, studying their evolution and possible function only became feasible when CGC1 made them visible.

Finally, the CGC1 genome should enable synthetic biology in what is arguably the best understood and most experimentally tractable model metazoan, *C. elegans* (Corsi et al. 2015). The combination of powerful computational tools for predicting gene functions, efficient methods for large-scale DNA synthesis, and effective protocols for targeted genome mutagenesis (e.g., CRISPR) has made it practical to analyze a genome, devise a model for how that genome could be functionally modified (e.g., by streamlining or recoding it), synthesize an artificial revision of that genome, propagate the revised genome in a living cell, and experimentally test whether this confirms the functional model (Venter et al. 2022; Hoose et al. 2023). For such work, it is necessary (though not sufficient) that the organism's genome assembly be entirely complete and accurate. Partly for that reason, attempts at such synthetic biology have focused heavily until now on modifying the genomes of single-celled organisms, such as the bacterium *Escherichia coli* or the baker's yeast *S. cerevisiae* (Pelletier et al. 2021; Zhao et al. 2023; Tian et al. 2024). An ultimate goal of synthetic biology is to scale this work upward to the most complex multicellular eukaryotes of interest, such as humans and other mammals or such as the crop plants that sustain human life (Boeke et al. 2016; Tuncel et al. 2023). However, genome engineering faces a wide gap in difficulty between *E. coli* and *Homo sapiens*. *C. elegans* was selected as a model organism in part because its complexity is greater than that of a microbe but is less than that of a human. This intermediate complexity may make *C. elegans* an ideal test system for synthetic biology of animal genomes and may make CGC1 the first *C. elegans* genome fully suited for such synthetic biology.

Methods

Orienting and visualizing genome assemblies with associated data

In this paper, we use the terms "left end" and "right end" of the *C. elegans* CGC1 reference chromosomes as designated in the original genetic map of Brenner (1974); the strands of CGC1 follow the original strand choice of the *C. elegans* N2 assembly (The *C. elegans* Sequencing Consortium 1998). We drew dot plots between sequences of two chromosomes or chromosome sets with GenomeMatcher 3.0.6 (Ohtsubo et al. 2008) using the NUCmer program in MUMmer 4.0.0rc1 (Kurtz et al. 2004); we drew dot plots between sequences of two tandem repeat regions with Gepard 2.1 (Krumsiek et al. 2007). We visualized multiple alignment of reads and read coverage of the CGC1 genome with the Integrative Genomics Viewer (IGV) 2.12.3 (IGV) (Robinson et al. 2011). We visualized gene structures, StringTie2 transcripts, and other genome annotations with JBrowse2 (Diesh et al. 2023).

C. elegans strain nomenclature

PD1074 (Yoshimura et al. 2019) is the original strain name of CGC1, which we renamed for two reasons: to have a strain name exactly identical to our telomere-to-telomere gap-free assembly name and to have a strain/assembly name that would be easy for biologists to remember and use. CGC1 (i.e., PD1074) was derived from VC2010 by selfing an individual hermaphrodite and allowing the population to expand for approximately 10 generations before expansion and mass-freezing at the CGC stock center. VC2010 was, in turn, a Moerman Gene Knockout Laboratory subculture of N2 obtained from the *Caenorhabditis* Genetics Center (https://cgc.umn.edu) in 2002 and first described by Flibotte et al. (2010).

C. elegans genome and gene annotation versions

In analyzing gDNA from the N2 assembly, we have used the most recent stable version from WormBase, which has remained unchanged from WormBase release WS235 (made available by FTP on November 28, 2012) to at least WormBase release WS294 (made available by FTP on August 16, 2024; for full details of WormBase version releases, see https://wormbase.org/about/release_schedule). A specific copy of the N2 genome sequence that we used was downloaded from WormBase WS289 (Supplemental Table S19), but any version from WS235 to WS294 would have been identical. In analyzing annotations from the N2 assembly (either for genes or for repetitive gDNA regions) (Supplemental Table S19), we used WormBase release WS292 (made available by FTP on February 29, 2024). On one hand, the gene annotations in WS292 will not be exactly the same as those in any other version of WormBase, so it might superficially seem that our analyses only apply to N2 genes from WS292 alone. However, gene structures from WormBase have been manually curated for up to 25 years; by 2024, the gene annotations for N2 had become very heavily reviewed and quite stable, with relatively minor changes in overall gene content over time. For this reason, in the main text, we generally refer to "N2 genes" rather than "WS292 genes." It is their origin from N2 that is most relevant to biologists, being a specific snapshot from WS292 is technically important but not of central biological relevance to our analyses here.

Sample isolation

For PacBio and Nanopore DNA sequencing, *C. elegans* of the strain CGC1 was grown on enriched nematode growth medium (NGM) plates seeded with an *E. coli* OP50 bacterial lawn (Brenner 1974). Roughly synchronized adult animals were obtained by standard hypochlorite synchronization. The animals were washed off agar plates with cold buffer (50 mM NaCl, 5 mM EDTA at pH 7.5) into 15 mL conical tubes. Worms were centrifuged at 500*g* for 3 min using a swinging bucket rotor in a tabletop centrifuge (Eppendorf Centrifuge 5702R). To decrease bacterial populations, pelleted animals were resuspended in 1 mL of cold buffer (50 mM NaCl, 5 mM EDTA at pH 7.5) and swirled persistently. Using a pipette, we gently transferred the worms to the top of a new tube with a 10 mL cold sucrose cushion (160 mM sucrose,

50 mM NaCl, 5 mM EDTA at pH 7.5) and centrifuged at 100*g* for 2 min in a tabletop centrifuge (Eppendorf Centrifuge 5702R), and the supernatant including bacteria was removed. Worm pellets at the bottom were used for further gDNA isolation.

Genome assembly and refinement of the CGC1 genome

We first assembled HiFi reads into 602 contigs with HiCanu 2.1.1 (Nurk et al. 2020), consolidated them to 80 nonredundant contigs (Supplemental Table S2A) with purge_dups 1.2.5 (Guan et al. 2020), and further reduced them to 61 nonredundant contigs by manually inspecting their order after aligning them to the VC2010 assembly with minimap2 v2.13 (Li 2018; Supplemental Fig. S1). This left us with 54 gaps between neighboring contigs, all of which could be spanned by Nanopore ultralong reads. Eleven gaps were <1 kb and could be determined by the consensus of Nanopore ultralong reads. However, the remaining 43 gaps were filled with long tandem repeats ≥10 kb in size, making their correct sequence more difficult to determine. We therefore assembled these 43 gaps manually, first using Nanopore ultralong reads to close gaps throughout the genome assembly and then using Nanopore and HiFi reads to correct errors and generate the complete genome (Fig. 1).

Comparison of CGC1 assembly with automated assemblies with Nanopore R10 data

To compare our CGC1 assembly process and product with the products of recent experimental and computational tools (Nanopore R10 and automated assemblers hifiasm [Cheng et al. 2021] and Verkko [Rautiainen et al. 2023]), we obtained Nanopore chemistry R10 read data for the population of animals studied here and performed assembles using these software tools.

The CGC1 assembly utilized a semiautomatic approach of filling gaps spanned by Nanopore ultralong reads (Fig. 1), with the semiautomated approach chosen because we were concerned that it would be difficult to fill the gaps using automated assembly methods. For comparison, we used the two genome assemblers hifiasm 0.19.7 and Verkko 2.2.1 (Antipov et al. 2025) to assemble R9 and R10 read data sets (Supplemental Table S1) in addition to PacBio HiFi reads to output assembled contigs, and Supplemental Table S2B shows the statistics of individual assembly. Specifically, from the two assemblers, hifiasm and Verkko, and the two chemistry Nanopore data sets, R9 and R10 (Supplemental Table S2B), we generated four assemblies: hifiasm/R9, Verkko/R9, hifiasm/R10, and Verkko/R10. From an overall perspective, each of the four assemblies appears to be consistent with the CGC1 assembly because each covered individual chromosomes of the CGC1 assembly with one to three assembled contigs (Supplemental Table S2B).

At the nucleotide level, we examined all of 106 positions at which the CGC1 assembly differed from all of the four assemblies generated from the four combinations of hifiasm/Verkko and R9/R10. To test whether the nucleotide at each position matched the CGC1 assembly or not, care had to be taken to align reads to a duplicated region (CHROMOSOME_V: 9,194,944–9,278,986) containing two repeated ∼42 kb subregions. To span these duplicated subregions, we used Nanopore ultralong reads >50 kb in length while avoiding shorter PacBio HiFi reads (and subreads) ∼20 kb in length. However, our examination required excluding 40 positions for which Nanopore read alignments were unreliable because these positions had very low read coverage, were in the long-duplicated regions, or were in short tandem repeat regions of CT repeats.

The remaining 66 positions still had 35 positions in the long-duplicated region. We then tested whether the nucleotide at each position matched the corresponding nucleotide in the CGC1 assembly (or nucleotide X in all of the other four assemblies, respectively) by checking the stringent condition that the proportion of Nanopore reads matching CGC1 (X) was >90% and 50% greater than the proportion of reads matching X (CGC1). For example, "A" at CHROMOSOME_X: 5,200,868 was replaced by "T" in the four assemblies, and the focal "A" matched 100% of 290 Nanopore reads, whereas "T" matched only 1.7% of 289 reads. We observed that 29 of the 66 positions fulfilled the above stringent criteria. Of these, four matched the CGC1 assembly, whereas 25 instead matched the four assemblies, indicating that the CGC1 assembly should be corrected to agree with them. Those 25 positions comprised only 0.000024% (2.4×10^{-7}) of the whole CGC1 assembly. The long-duplicated region contained a substantial fraction of the discrepancies; for example, 19 of the 25 corrected positions were in this region, complicating current assembly algorithms and increasing the likelihood of generating false-positive SNVs. We also examined the increase in the number of positions when we relaxed the stringent criteria for considering nucleotide site data; this reduced the discrepancy from 50% to 2%, but the increase was smaller, and nine and 33 positions matched the CGC1 and four assemblies, respectively.

Each of the four hifiasm/Verkko assemblies failed to assemble complex tandem repeat regions, divided the 45S rDNA array into several smaller fragments (Supplemental Fig. S6), and output a single contig for pSX1 that was much shorter than the pSX1 region ∼153 kb in length in the CGC1 assembly (Supplemental Table S2B; Supplemental Fig. S7).

Identifying tandem repeats

We used mTR to identify tandem repeats within CGC1 and other *C. elegans* genome assemblies (Morishita et al. 2021). Tandem repeats are usually classified as short tandem repeats with a repeat unit length of ≤6 nt and minisatellites with a repeat unit length of ≥6 nt.

Filling gaps in five of 43 complex regions

This HiFi error correction step was successful in correcting errors in 0.014%–0.024% of sequence in three of the five complex regions (Table 2). However, in two other regions (Supplemental Fig. S3), polishing the Nanopore consensus by HiFi reads increased a number of errors because the consensus at an erroneous base did not account for most bases, and more than one base could be prevalent in the column. This suggested that although HiFi reads are free from sequence bias in nonrepetitive regions (Wenger et al. 2019), they may have sequence bias in tandem repeats. For example, because an adenine in a 27-mer unit on CHROMOSOME_X (nt 5,581,121–5,673,280) was often called as a cytosine erroneously in HiFi reads (Supplemental Fig. S3E), mapping HiFi reads to the Nanopore consensus caused an error rate of ∼4% (1/27). To correct for this error, we reasoned that HiFi reads are obtained as a consensus of subreads, that sequencing errors of subreads occur at random, and that mapping PacBio subreads instead of HiFi reads might help us determine which consensus was correct. Mapping PacBio subreads, we validated and chose the Nanopore consensus as the reference sequence (Supplemental Fig. S3D,E).

Identifying variants in tandemly repeated sequence elements

It is essential to ensure that tools used for assembly of tandemly repeated regions do not suffer from the artifact of regression to the mean, which would assign variants present in a minority of repeats to a majority sequence. Conversely, an approach that starts with identifying SNVs and indels that are characteristic of a minority of repeats and requiring a perfect alignment of these provides an

opportunity for a definitive long-range assembly of such regions. Once a draft assembly has been produced through such an approach, a next challenge entails an assessment of reliability. One approach here would be to search a set of long reads (for ultralong reads that span the pSX1 array, see Supplemental Fig. S16) for perfect matches (or, closest partial matches) of the variant repeat; however, this process is confounded in cases in which the error rate is much higher than the variant rate. Nanopore's ultralong reads have a sequence error rate of ~5%; so, for example, ~8 nt of the 172 nt pSX1 will on average be incorrect. We instead attempted to locate SNVs and indels unique to the variant and listed perfect matches of a short motif string around a focal SNV (or indel) because a shorter string is less likely to include sequencing errors. Indeed, the occurrences of the short motif in two different Nanopore reads showed a clear pattern of reproducibility in the occurrences of the two variants in the assembled pSX1 array (Supplemental Figs. S17–S21).

Analysis and error correction of the 45S rDNA array

Figure 5 shows how the 45S rDNA array was assembled from Nanopore reads. To do this reliably with Nanopore reads that had an error rate of ~5% for the six SNVs, we searched for positions that perfectly matched 11 bases centered on each of the six SNV sites. Longer strings are more convenient for selecting unique positions but are less likely to match the corresponding positions owing to sequencing errors. In contrast, shorter strings are less likely to introduce sequencing errors but may yield a larger number of positions. We selected 11 matching bases from candidate $(2k+1)$ matching bases for $k=3$, 4, 5, 6, and 7 by considering the tradeoff between the expected number of positions and the probability of perfect matches with a sequencing error rate. Precisely, to select positions with 11 matching bases, we consider a tradeoff for a $(2k+1)$-mer between the expected number of positions, N, at which the $(2k+1)$-mer matches a random $(2k+1)$-mer in the worm genome of size ~100 Mb, and the probability, P, that the $(2k+1)$-mer matches the original position with a 5% sequencing error rate. For $k=3$, 4, 5, 6, and 7, the values of N are 6103.5, 381.5, 23.8, 1.5, and 0.1, and P is 0.70, 0.63, 0.57, 0.51, and 0.46. Considering the trade-off, we set k to five and selected 11-mer.

The size of the assembled 45S rDNA array reached to 772 kb, with confirmation from an independent data set consisting of PacBio HiFi reads. Notable here, of the six SNV positions, SNVs at four positions (nt 1264, 2070, 4161, and 6018) occurred only once in the assembly. The minor nucleotides of these positions were covered by 60, 62, 64, and 60 HiFi reads, respectively, whereas the reference (majority) nucleotides were covered by 6614, 6619, 6727, and 6584 reads (Fig. 5A). This is consistent with an array in which one unit has the minor nucleotide and about 110 units (e.g., 6614/60) have a representative nucleotide within the 45S rDNA array assembly. Thus, the total length of the array is ~799 kb (111 units times 7197 nt per unit), which is close to 772 kb, the size of the assembled array. Another study estimated copy numbers of 98 to 114 units from the Nanopore reads of strain N2 (Ding et al. 2022).

Figure 5D shows the read coverage distribution in the final consensus, and deletions were observed at 541 positions, all of which were homopolymers (strings of identical nucleotides) and likely generated from Nanopore sequencing errors. At 587 positions of the Nanopore consensus, >20% of Nanopore reads had mismatches that did not match the consensus bases. This discrepancy is likely owing to bias in Nanopore sequencing because 13 bases centered on each mismatch did not match any PacBio HiFi reads, whereas 13 bases centered on the consensus base matched averaged 6633 locations in PacBio HiFi reads. Overall, the final consensus had 107 copies of the 45S rDNA representative repeat unit with 30 SNVs and three insertions (for illustration, see Fig. 5B).

General manipulations of genomic data

When possible, we used mamba to install and run version-controlled software environments from Bioconda (Grüning et al. 2018). For reformatting or parsing of computational results, we used Perl scripts either developed for general use or custom-coded for a given analysis. All such Perl scripts (named below with italics and the suffix "*.pl*") were archived on GitHub (https://github.com/schwarzem/ems_perl). Internet sources (URLs) for other software are listed in Supplemental Table S20.

We extracted gene-encoded DNA or protein sequences from gDNA sequences and gene annotations in GTF/GFF format with gffread 0.12.7 (Pertea and Pertea 2020) using the arguments "*-g [genomic sequence FASTA] -o /dev/null -C --keep-genes -P -V -H -y [translated protein sequence] -w [transcribed DNA sequence].*" We identified overlaps between different features of the CGC1 assembly with the *intersect* tool of BEDTools 2.30.0 (Quinlan and Hall 2010) using the argument "*-loj*." Before testing them for overlaps, genomic features were converted from GTF/GFF to BED format with *gff2bed* from BEDOPS 2.4.41 (Neph et al. 2012) using the argument "*--do-not-sort.*"

Obtaining published nematode genomic data

Published genome sequences, coding sequences, proteomes, and gene annotations were downloaded from WormBase (Sternberg et al. 2024), WormBase ParaSite (Howe et al. 2017), or the Open Science Framework (Supplemental Table S19; Foster and Deardorff 2017). Published ribodepleted RNA-seq data of *C. elegans* (Wang et al. 2022) were downloaded from the European Nucleotide Archive (Supplemental Table S19).

Identifying shared and unique nucleotide sequences in the N2 and CGC1 assemblies

To determine nucleotide coordinates for regions of the CGC1 assembly that either were uniquely shared by the N2 assembly or were not so shared, we first generated an alignment of uniquely matching sequences of the two assemblies with NUCmer using the arguments "*--mum --breaklen 1 --maxgap 0 --batch 1*" and then filtered their delta alignment with dnadiff, both NUCmer and dnadiff being from MUMmer4 4.0.0rc1 (Marçais et al. 2018). We converted the resulting filtered delta alignment into MAF format with delta2maf from Mugsy 1.2.3 (Angiuoli and Salzberg 2011) and then from MAF to GFF format with maf-convert from LAST 1548 (Kielbasa et al. 2011) using the argument "*gff.*" Readable genomic nucleotide coordinate annotations were added to the GFF alignment with *id2nucmer_gffs_14jul2024.pl*. The GFF alignment was converted to BED format with gff2bed from BEDOPS 2.4.41 (Neph et al. 2012) using the argument "*--do-not-sort*," sorted with *sort* from Linux using the arguments "-k1,1 -k2,2n," merged into nonredundant genomic regions with BEDTools merge from BEDTools 2.30.0 using the argument "*-delim* "|"," and converted back to GFF format with *convert_bed2gff.pl* from AGAT 1.3.3 (https://zenodo.org/records/3549547). Source and sequence type annotations in the GFF alignment were converted from "data\tgene" to "maf-convert\tregion" with inline Perl. Alignment regions were given specific sequence-block names with *id2nucmer_gffs_ 26aug2024.pl*. The annotated GFF alignment was converted back to BAM format with gff2bed from BEDOPS 2.4.41 using the argument "*--do-not-sort*" and was genomically

sorted with BEDTools sort from BEDTools 2.30.0 using the argument "*-g [CGC1 genome assembly].*" Coordinates of the CGC1 genome that were not part of the resulting N2 alignment were extracted into a BED file with BEDTools complement from BEDTools 2.30.0 using the arguments "*-i [N2-CGC1 genome alignment] -g [CGC1 genome assembly].*"

Mapping N2 genetic content to the CGCI assembly

We obtained canonical N2 gene annotations from WormBase WS292. Before mapping to them to CGC1, we split them into separate GFF3 annotation files with protein-coding genes, pseudogenes, and ncRNAs by selecting for the biotypes "protein_coding," "pseudogene," or any other biotypes (respectively); this first allowed us to use different mapping parameters when appropriate for different gene types and later allowed us to compare de novo protein-coding genes separately to N2 protein-coding genes versus N2 pseudogenes. We mapped all three sets of canonical N2 WS292 gene annotations to the CGC1 assembly with Liftoff 1.6.3 (Shumate and Salzberg 2021) using the arguments "*-s 0.9 -cds -polish -copies -chroms [table of identical chromosomes].*" For mapping of ncRNA genes, which were otherwise too small to map reliably, we also used the argument "*-flank 1.0.*"

Prediction of protein-coding genes

We predicted protein-coding genes in the CGC1 assembly de novo with AUGUSTUS 3.5.0 (Stanke et al. 2008) using the arguments "*--strand=both --genemodel=partial --noInFrameStop=true --singlestrand=false --maxtracks=3 --alternatives-from-sampling=true --alternatives-from-evidence=true --minexonintronprob=0.1 --minmeanexonintronprob=0.4 --uniqueGeneId=true --protein=on --introns=on --start=on --stop=on --cds=on --codingseq=on --UTR=off --species=caenorhabditis --extrinsicCfgFile=[augustus_dir]/config/extrinsic/extrinsic.ME.cfg --progress=true --gff3=on --outfile=[gene_predictions].aug --hintsfile=[hints].gff.*" To generate hints for AUGUSTUS, we merged CDS DNA sequences both for N2 protein-coding genes from WormBase WS289 and for VC2010 v.1 protein-coding genes from WormBase WS268; we then mapped the CDS DNA sequences to CGC1 gDNA with BLAT v362 (Kent 2002) using the argument "*-minIdentity=92,*" filtered the BLAT alignments with pslCDnaFilter with the argument "*-maxAligns=1,*" and converted the alignments from PSL to GFF format with *blat2hints.pl* from AUGUSTUS.

StringTie2 prediction of transcripts and transcribed genomic loci in CGCI

We predicted transcription units in the CGC1 assembly using paired-end rRNA-depleted *C. elegans* RNA-seq data from six N2 and *lpd-3* replicates (Wang et al. 2022) with StringTie2 (version 2.2.1; Kovaka et al. 2019) using the argument "-G *[AUGUSTUS gene predictions]*" to identify transcripts corresponding with de novo protein-coding gene predictions (and, by exclusion, identify possible ncRNA transcripts). Before running StringTie2, we aggregated the RNA-seq reads into a single pair of read files; we used HISAT2 2.2.1 (Kim et al. 2019) to index the CGC1 genome with "*hisat2-build*" and to map pooled RNA-seq reads to it with "*hisat2 --summary-file [mapping summary] -x [CGC1 HISAT2 index] -1 [read-pair end 1] -2 [read-pair end 2] -S [alignment in SAM format]*"; we then reformatted the resulting SAM alignment to a sorted, indexed BAM alignment with "*view –b,*" "*sort,*" and "*index*" in SAMtools 1.17 (Danecek et al. 2021).

Long RNA-seq analysis of TRs

We mapped long RNA-seq reads to the CGC1 assembly with winnowmap2 (Jain et al. 2022). Of 174 tandem repeat regions of ≥5 kb in CGC1, 10 regions had alignments; however, six alignments showed no splicing of the long RNA-seq reads, suggesting that they might arise from DNA contamination rather than mRNA. Four other alignments with splicing were correlated with gene structures via JBrowse2 to distinguish likely intronic transcripts from possible ncRNAs.

Annotation of gene products

For protein-coding genes, we predicted both N-terminal signal sequences and transmembrane alpha-helical anchors with Phobius 1.01 (Käll et al. 2004), reformatting results with *tabulate_phobius_hits.pl*. We predicted coiled-coil domains with Ncoils 2002.08.22 (Lupas 1996), reformatting results with *tabulate_ncoils_x.fa.pl*. We predicted low-complexity regions with PSEG 1999.06.10 (Wootton 1994) using the argument "*-l*" and reformatting results with *summarize_psegs.pl*. We identified protein domains from the Pfam 37.0 database (Mistry et al. 2021) with hmmscan in HMMER 3.3.2, using the arguments "*--cut_ga*" to impose family-specific significance thresholds, "*-o /dev/null*" to discard text outputs, and "*--tblout*" to export tabular outputs; Pfam results were reformatted with *pfam_hmmscan2annot.pl*. We also identified protein domains from the InterProScan 5.57–90.0 database with *interproscan.sh* (Paysan-Lafosse et al. 2023) using the arguments "*-dp iprlookup goterms*" and reformatting results with *tabulate_iprscan_tsv.pl*. We identified orthologs between protein-coding gene sets from *C. elegans* (N2 and CGC1), *C. nigoni*, *C. remanei*, and *C. elegans* N2 transposon-encoded proteins with OrthoFinder 2.5.4 (Emms and Kelly 2019) using the arguments "*-S diamond_ultra_sens –og*"; results were reformatted with *prot2gene_ofind.pl* and *genes2omcls.pl*. For all annotation analyses except OrthoFinder, full proteomes were used; for OrthoFinder, we used maximum-isoform proteome subsets generated with *get_largest_isoforms.pl*. An overall annotation table was constructed from these and other gene annotations (such as overlaps between lifted-over N2 and CGC1 AUGUSTUS genes) with *add_tab_annots.pl*.

For StringTie2 transcripts, we identified ncRNA motifs from the Rfam 14.10 database (Kalvari et al. 2021) with cmsearch in INFERNAL 1.1.5 (Nawrocki and Eddy 2013) using the arguments "*--cut_ga*" to impose family-specific significance thresholds, "*-o /dev/null*" to discard text outputs, and "*--tblout*" to export tabular outputs.

Detection of proteomic spectra matching predicted *C. elegans* proteins

We downloaded mass spectrometric data in RAW file format (Supplemental Table S19) from the PRIDE database (Perez-Riverol et al. 2022) for three surveys of the *C. elegans* proteome (Xia et al. 2018; Müller et al. 2020; Ceron-Noriega et al. 2023). We reformatted these data files from RAW to mzML format with thermorawfileparser 1.4.4 (Hulstaert et al. 2020), using the arguments "*--format 2 --metadata 1.*" We mapped protein spectra in these files to *C. elegans* protein sequences (from either our de novo AUGUSTUS gene predictions in CGC1 or from the preliftover N2 protein-coding genes from WormBase WS292) with Comet (Eng et al. 2013) from Crux 4.1 (McIlwain et al. 2014) using the arguments "*--nucleotide_reading_frame 0 --decoy_search 1.*" We statistically analyzed these three mappings as a single collective data set with Percolator (Käll et al. 2007) from Crux 4.1 using the argument "*--decoy-prefix decoy_.*" We classified protein-coding genes (from

either AUGUSTUS predictions from the N2 or CGC1 assembly) as being detected in protein spectra if they showed mapped spectra with a statistical significance (corrected for multiple hypothesis testing) of $q \leq 0.01$. This threshold was chosen to limit the rate of false positives to 1%; it is necessarily arbitrary and will by nature entail some false negatives as well as false positives.

Observation of ART2/RRT15 homologs in eukaryotes

We detected homologs of putative *C. elegans* ART2/RRT15-domain encoding genes by examining proteins associated with their shared InterPro motif (https://www.ebi.ac.uk/interpro/entry/InterPro/IPR052997) and by conducting BLASTP searches of the NCBI nonredundant protein database (nr; https://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastp). Partial overlap of the ART2/RRT15-domain encoding genes with the previously observed *C. elegans* ncRNA gene F31C3.14 (Lu et al. 2011) was detected by BLASTN of the *C. elegans* reference genome (https://wormbase.org/tools/blast_blat) and confirmed by StringTie2 transcript overlaps (Supplemental Table S13).

Disqualifying novel AUGUSTUS gene predictions or StringTie2 transcript loci

Most novel AUGUSTUS predictions (Supplemental Table S12) were identified as those that did not overlap with lifted-over N2 protein-coding genes and that at least partially overlapped with sequences specific to the CGC1 assembly (i.e., that were not shared with the N2 assembly). For AUGUSTUS predictions that at least partially overlapped sequences specific to the CGC1 assembly, we disqualified them if they met any of the following criteria: having orthology to a transposon-encoded protein, consisting of ≥50% low-complexity residues, or encoding 40 or more transmembrane domains. For other AUGUSTUS predictions that overlapped only N2-shared genomic sequences but that still failed to overlap N2 protein-coding genes, we disqualified them both by the above criteria and also by the following criteria: overlapping with an annotated N2 pseudogene, overlapping with a repetitive DNA region annotated with "transposon," or encoding a protein with the transposon-associated InterPro domain "Transposase, Tc5, C-terminal [IPR007350]."

For StringTie2 transcript loci, we identified those that at least partially overlapped sequences specific to the CGC1 assembly and then disqualified them as follows: We excluded any overlap with any lifted-over N2 protein-coding gene; we excluded any that overlapped with a AUGUSTUS-predicted protein-coding gene; we excluded any that overlapped with a lifted-over N2 ncRNA gene; and we excluded any transcripts encoding rRNA motifs from Rfam. For StringTie2 transcript loci that overlapped only N2-shared gDNA, we disqualified them by the same criteria but then also excluded any overlap with a lifted-over N2 pseudogene.

We used several criteria to identify which of the 2006 new protein-coding genes predicted by AUGUSTUS represented real *C. elegans* genes. Out of the 227 that overlapped CGC1-specific gDNA, we discounted 14 whose protein products showed orthology to *C. elegans* transposon-encoded proteins. We also discounted 30 genes likely to be mispredictions from repetitive gDNA: 28 whose proteins consisted of ≥50% low-complexity sequence and two whose proteins were predicted to encode 40 or 60 transmembrane domains.

Testing unmapped N2 genes for StringTie2 matches

Coding DNA sequences (CDS DNAs) for 75 unmapped N2 genes (Table 3; Supplemental Table S11) were extracted from the N2 genome with GFFread as described above. A BLASTN-searchable database was generated from StringTie2 sequences with makeblastdb from BLAST 2.14.0 ref. The StringTie2 DNA database was then searched for matches to the unmapped N2 CDS DNAs with BLASTN from BLAST 2.14.0. Results were examined by eye to identify full-length 100%-identical matches of CDSs to StringTie2 transcripts.

Data access

All raw sequencing reads and genome assembly for *C. elegans* CGC1 generated in this study have been submitted to DDBJ (https://www.ddbj.nig.ac.jp/) under accession number PRJDB19205 (Biosamples SAMD00841453-SAMD00841456). Predicted genes and transcripts have been archived in the Open Science Framework (https://osf.io/n3fdy). We anticipate that gene predictions will also be added to GenBank after they have been curated by WormBase. Supplemental Perl scripts are available in the Supplemental Code.

Competing interest statement

The authors declare no competing interests.

Acknowledgments

We thank Jun Yoshimura for generating a previous VC2010 assembly, Yuta Suzuki for basecalling reads, and the Information Technology Center, The University of Tokyo, for permission to use their supercomputing facilities (Oakbridge-CX and mdx). This work was supported by the Japan Agency for Medical Research and Development (AMED) 24tm0424219h0004 and Japan Society for the Promotion of Science KAKENHI grant number JP22H04925 (PAGS) to S.M., an Arnold O. Beckman Postdoctoral Independence Award and American Heart Association Postdoctoral Fellowship to M.J.S., National Institutes of Health (NIH) grant R35GM130366 to A.Z.F., an NIH resource grant P40OD010440 for the *Caenorhabditis* Genetics Center to A.E.R., and Cornell institutional funds to E.M.S.

Author contributions: S.M., A.Z.F., and E.M.S. conceived and supervised the study and had primary responsibility for writing the manuscript with input from all authors. K.I., S.M., and Y.S. performed genome assembly and computational analyses of genomic data. E.M.S. systematized and executed gene mapping and annotation. M.J.S., K.L.A., D.-E.J., M.K., Y.T., Y.I., A.E.R., and L.W. carried out experiments with *C. elegans* and contributed to the analysis of genome dynamics and transcription products. C.O. and H.K. sequenced the DNA materials. All authors have approved the manuscript.

References

Alkan C, Sajjadian S, Eichler EE. 2011. Limitations of next-generation genome sequence assembly. *Nat Methods* **8:** 61–65. doi:10.1038/nmeth.1527

Angiuoli SV, Salzberg SL. 2011. Mugsy: fast multiple alignment of closely related whole genomes. *Bioinformatics* **27:** 334–342. doi:10.1093/bioinformatics/btq665

Antipov D, Rautiainen M, Nurk S, Walenz BP, Solar SJ, Phillippy AM, Koren S. 2025. Verkko2 integrates proximity-ligation data with long-read De Bruijn graphs for efficient telomere-to-telomere genome assembly, phasing, and scaffolding. *Genome Res* **35:** 1583–1594. doi:10.1101/gr.280383.124

Awad M, Gan X. 2023. GALA: a computational framework for *de novo* chromosome-by-chromosome assembly with long reads. *Nat Commun* **14:** 204. doi:10.1038/s41467-022-35670-y

Biswas S, Gurdziel K, Meller VH. 2024. siRNA that participates in *Drosophila* dosage compensation is produced by many 1.688X and 359 bp repeats. *Genetics* **227:** iyae074. doi:10.1093/genetics/iyae074

Boeke JD, Church G, Hessel A, Kelley NJ, Arkin A, Cai Y, Carlson R, Chakravarti A, Cornish VW, Holt L, et al. 2016. Genome engineering: the Genome Project-Write. *Science* **353:** 126–127. doi:10.1126/science.aaf6850

Brenner S. 1974. The genetics of *Caenorhabditis elegans*. *Genetics* **77:** 71–94. doi:10.1093/genetics/77.1.71

Bush ZD, Naftaly AFS, Dinwiddie D, Albers C, Hillers KJ, Libuda DE. 2025. Transposable elements and heterochromatic regions are enriched for structural variation and sequence divergence in the genome of wild-type *Caenorhabditis elegans*. *G3 (Bethesda)* jkaf092. doi:10.1093/g3journal/jkaf092

Carlton PM, Davis RE, Ahmed S. 2022. Nematode chromosomes. *Genetics* **221:** iyac014. doi:10.1093/genetics/iyac014

The *C. elegans* Sequencing Consortium. 1998. Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science* **282:** 2012–2018. doi:10.1126/science.282.5396.2012

Ceron-Noriega A, Almeida MV, Levin M, Butter F. 2023. Nematode gene annotation by machine-learning-assisted proteotranscriptomics enables proteome-wide evolutionary analysis. *Genome Res* **33:** 112–128. doi:10.1101/gr.277070.122

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

Corsi AK, Wightman B, Chalfie M. 2015. A transparent window into biology: a primer on *Caenorhabditis elegans*. *Genetics* **200:** 387–407. doi:10.1534/genetics.115.176099

Crombie TA, Zdraljevic S, Cook DE, Tanny RE, Brady SC, Wang Y, Evans KS, Hahnel S, Lee D, Rodriguez BC, Zhang G, et al. 2019. Deep sampling of Hawaiian *Caenorhabditis elegans* reveals high genetic diversity and admixture with global populations. *eLife* **8:** e50465. doi:10.7554/eLife.50465

Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T, McCarthy SA, Davies RM, et al. 2021. Twelve years of SAMtools and BCFtools. *GigaScience* **10:** giab008. doi:10.1093/gigascience/giab008

Denton JF, Lugo-Martinez J, Tucker AE, Schrider DR, Warren WC, Hahn MW. 2014. Extensive error in the number of genes inferred from draft genome assemblies. *PLoS Comput Biol* **10:** e1003998. doi:10.1371/journal.pcbi.1003998

Diesh C, Stevens GJ, Xie P, De Jesus Martinez T, Hershberg EA, Leung A, Guo E, Dider S, Zhang J, Bridge C, et al. 2023. JBrowse 2: a modular genome browser with views of synteny and structural variation. *Genome Biol* **24:** 74. doi:10.1186/s13059-023-02914-z

Ding Q, Li R, Ren X, Chan LY, Ho VWS, Xie D, Ye P, Zhao Z. 2022. Genomic architecture of 5S rDNA cluster and its variations within and between species. *BMC Genomics* **23:** 238. doi:10.1186/s12864-022-08476-x

Doitsidou M, Jarriault S, Poole RJ. 2016. Next-generation sequencing-based approaches for mutation mapping and identification in *Caenorhabditis elegans*. *Genetics* **204:** 451–474. doi:10.1534/genetics.115.186197

Ellis RE, Sulston JE, Coulson AR. 1986. The rDNA of *C. elegans*: sequence and structure. *Nucleic Acids Res* **14:** 2345–2364. doi:10.1093/nar/14.5.2345

Emms DM, Kelly S. 2019. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol* **20:** 238. doi:10.1186/s13059-019-1832-y

Eng JK, Jahan TA, Hoopmann MR. 2013. Comet: an open-source MS/MS sequence database search tool. *Proteomics* **13:** 22–24. doi:10.1002/pmic.201200439

Essers PB, Nonnekens J, Goos YJ, Betist MC, Viester MD, Mossink B, Lansu N, Korswagen HC, Jelier R, Brenkman AB, et al. 2015. A long noncoding RNA on the ribosome is required for lifespan extension. *Cell Rep* **10:** 339–345. doi:10.1016/j.celrep.2014.12.029

Flibotte S, Edgley ML, Chaudhry I, Taylor J, Neil SE, Rogula A, Zapf R, Hirst M, Butterfield Y, Jones SJ, et al. 2010. Whole-genome profiling of mutagenesis in *Caenorhabditis elegans*. *Genetics* **185:** 431–441. doi:10.1534/genetics.110.116616

Foster ED, Deardorff A. 2017. Open science framework (OSF). *J Med Libr Assoc* **105:** 203–206. doi:10.5195/jmla.2017.88

Gems D, Riddle DL. 2000. Defining wild-type life span in *Caenorhabditis elegans*. *J Gerontol A Biol Sci Med Sci* **55:** B215–B219. doi:10.1093/gerona/55.5.B215

Grüning B, Dale R, Sjodin A, Chapman BA, Rowe J, Tomkins-Tinch CH, Valieris R, Köster J. 2018. Bioconda: sustainable and comprehensive software distribution for the life sciences. *Nat Methods* **15:** 475–476. doi:10.1038/s41592-018-0046-7

Guan D, McCarthy SA, Wood J, Howe K, Wang Y, Durbin R. 2020. Identifying and removing haplotypic duplication in primary genome assemblies. *Bioinformatics* **36:** 2896–2898. doi:10.1093/bioinformatics/btaa025

Hillier LW, Coulson A, Murray JI, Bao Z, Sulston JE, Waterston RH. 2005. Genomics in *C. elegans*: so many genes, such a little worm. *Genome Res* **15:** 1651–1660. doi:10.1101/gr.3729105

Hontz RD, Niederer RO, Johnson JM, Smith JS. 2009. Genetic identification of factors that modulate ribosomal DNA transcription in *Saccharomyces cerevisiae*. *Genetics* **182:** 105–119. doi:10.1534/genetics.108.100313

Hoose A, Vellacott R, Storch M, Freemont PS, Ryadnov MG. 2023. DNA synthesis technologies to close the gene writing gap. *Nat Rev Chem* **7:** 144–161. doi:10.1038/s41570-022-00456-9

Hori Y, Engel C, Kobayashi T. 2023. Regulation of ribosomal RNA gene copy number, transcription and nucleolus organization in eukaryotes. *Nat Rev Mol Cell Biol* **24:** 414–429. doi:10.1038/s41580-022-00573-9

Howe KL, Bolt BJ, Shafie M, Kersey P, Berriman M. 2017. Wormbase ParaSite: a comprehensive resource for helminth genomics. *Mol Biochem Parasitol* **215:** 2–10. doi:10.1016/j.molbiopara.2016.11.005

Hulstaert N, Shofstahl J, Sachsenberg T, Walzer M, Barsnes H, Martens L, Perez-Riverol Y. 2020. ThermoRawFileParser: modular, scalable, and cross-platform RAW file conversion. *J Proteome Res* **19:** 537–542. doi:10.1021/acs.jproteome.9b00328

Ishtayeh H, Achache H, Kroizer E, Rappaport Y, Itskovits E, Gingold H, Best C, Rechavi O, Tzur YB. 2021. Systematic analysis of long intergenic noncoding RNAs in *C. elegans* germline uncovers roles in somatic growth. *RNA Biol* **18:** 435–445. doi:10.1080/15476286.2020.1814549

Jain C, Rhie A, Hansen NF, Koren S, Phillippy AM. 2022. Long-read mapping to repetitive reference sequences using Winnowmap2. *Nat Methods* **19:** 705–710. doi:10.1038/s41592-022-01457-8

Johnson SM, Tan FJ, McCullough HL, Riordan DP, Fire AZ. 2006. Flexibility and constraint in the nucleosome core landscape of *Caenorhabditis elegans* chromatin. *Genome Res* **16:** 1505–1516. doi:10.1101/gr.5560806

Käll L, Krogh A, Sonnhammer EL. 2004. A combined transmembrane topology and signal peptide prediction method. *J Mol Biol* **338:** 1027–1036. doi:10.1016/j.jmb.2004.03.016

Käll L, Canterbury JD, Weston J, Noble WS, MacCoss MJ. 2007. Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nat Methods* **4:** 923–925. doi:10.1038/nmeth1113

Kalvari I, Nawrocki EP, Ontiveros-Palacios N, Argasinska J, Lamkiewicz K, Marz M, Griffiths-Jones S, Toffano-Nioche C, Gautheret D, Weinberg Z, et al. 2021. Rfam 14: expanded coverage of metagenomic, viral and microRNA families. *Nucleic Acids Res* **49:** D192–D200. doi:10.1093/nar/gkaa1047

Kent WJ. 2002. BLAT: the BLAST-like alignment tool. *Genome Res* **12:** 656–664. doi:10.1101/gr.229202

Kielbasa SM, Wan R, Sato K, Horton P, Frith MC. 2011. Adaptive seeds tame genomic sequence comparison. *Genome Res* **21:** 487–493. doi:10.1101/gr.113985.110

Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. 2019. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat Biotechnol* **37:** 907–915. doi:10.1038/s41587-019-0201-4

Kiontke KC, Herrera RA, Vuong E, Luo J, Schwarz EM, Fitch DHA, Portman DS. 2019. The long non-coding RNA *lep-5* promotes the juvenile-to-adult transition by destabilizing LIN-28. *Dev Cell* **49:** 542–555.e9. doi:10.1016/j.devcel.2019.03.003

Kobayashi T, Ganley AR. 2005. Recombination regulation by transcription-induced cohesin dissociation in rDNA repeats. *Science* **309:** 1581–1584. doi:10.1126/science.1116102

Kovaka S, Zimin AV, Pertea GM, Razaghi R, Salzberg SL, Pertea M. 2019. Transcriptome assembly from long-read RNA-seq alignments with StringTie2. *Genome Biol* **20:** 278. doi:10.1186/s13059-019-1910-1

Krumsiek J, Arnold R, Rattei T. 2007. Gepard: a rapid and sensitive tool for creating dotplots on genome scale. *Bioinformatics* **23:** 1026–1028. doi:10.1093/bioinformatics/btm039

Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg SL. 2004. Versatile and open software for comparing large genomes. *Genome Biol* **5:** R12. doi:10.1186/gb-2004-5-2-r12

Lee D, Zdraljevic S, Stevens L, Wang Y, Tanny RE, Crombie TA, Cook DE, Webster AK, Chirakar R, Baugh LR, et al. 2021. Balancing selection maintains hyper-divergent haplotypes in *Caenorhabditis elegans*. *Nat Ecol Evol* **5:** 794–807. doi:10.1038/s41559-021-01435-x

Lee H, Kim J, Lee J. 2023. Benchmarking datasets for assembly-based variant calling using high-fidelity long reads. *BMC Genomics* **24:** 148. doi:10.1186/s12864-023-09255-y

Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34:** 3094–3100. doi:10.1093/bioinformatics/bty191

Li R, Hsieh CL, Young A, Zhang Z, Ren X, Zhao Z. 2015. Illumina synthetic long read sequencing allows recovery of missing sequences even in the "finished" *C. elegans* genome. *Sci Rep* **5:** 10814. doi:10.1038/srep10814

Li R, Ren X, Ding Q, Bi Y, Xie D, Zhao Z. 2020. Direct full-length RNA sequencing reveals unexpected transcriptome complexity during *Caenorhabditis elegans* development. *Genome Res* **30:** 287–298. doi:10.1101/gr.251512.119

Lu ZJ, Yip KY, Wang G, Shou C, Hillier LW, Khurana E, Agarwal A, Auerbach R, Rozowsky J, Cheng C, et al. 2011. Prediction and characterization of noncoding RNAs in *C. elegans* by integrating conservation, secondary structure, and high-throughput sequencing and array data. *Genome Res* **21:** 276–285. doi:10.1101/gr.110189.110

Lupas A. 1996. Prediction and analysis of coiled-coil structures. *Meth Enzymol* **266:** 513–525. doi:10.1016/S0076-6879(96)66032-7

Marçais G, Delcher AL, Phillippy AM, Coston R, Salzberg SL, Zimin A. 2018. MUMmer4: a fast and versatile genome alignment system. *PLoS Comput Biol* **14:** e1005944. doi:10.1371/journal.pcbi.1005944

Maydan JS, Lorch A, Edgley ML, Flibotte S, Moerman DG. 2010. Copy number variation in the genomes of twelve natural isolates of *Caenorhabditis elegans*. *BMC Genomics* **11:** 62. doi:10.1186/1471-2164-11-62

McIlwain S, Tamura K, Kertesz-Farkas A, Grant CE, Diament B, Frewen B, Howbert JJ, Hoopmann MR, Käll L, Eng JK, et al. 2014. Crux: rapid open source protein tandem mass spectrometry analysis. *J Proteome Res* **13:** 4488–4491. doi:10.1021/pr500741y

Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, Sonnhammer ELL, Tosatto SCE, Paladin L, Raj S, Richardson LJ, et al. 2021. Pfam: the protein families database in 2021. *Nucleic Acids Res* **49:** D412–D419. doi:10.1093/nar/gkaa913

Morishita S, Ichikawa K, Myers EW. 2021. Finding long tandem repeats in long noisy reads. *Bioinformatics* **37:** 612–621. doi:10.1093/bioinformatics/btaa865

Müller JB, Geyer PE, Colaço AR, Treit PV, Strauss MT, Oroshi M, Doll S, Virreira Winter S, Bader JM, Köhler N, et al. 2020. The proteome landscape of the kingdoms of life. *Nature* **582:** 592–596. doi:10.1038/s41586-020-2402-x

Nawrocki EP, Eddy SR. 2013. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* **29:** 2933–2935. doi:10.1093/bioinformatics/btt509

Nelson DW, Honda BM. 1985. Genes coding for 5S ribosomal RNA of the nematode *Caenorhabditis elegans*. *Gene* **38:** 245–251. doi:10.1016/0378-1119(85)90224-0

Neph S, Kuehn MS, Reynolds AP, Haugen E, Thurman RE, Johnson AK, Rynes E, Maurano MT, Vierstra J, Thomas S, et al. 2012. BEDOPS: high-performance genomic feature operations. *Bioinformatics* **28:** 1919–1920. doi:10.1093/bioinformatics/bts277

Nurk S, Walenz BP, Rhie A, Vollger MR, Logsdon GA, Grothe R, Miga KH, Eichler EE, Phillippy AM, Koren S. 2020. HiCanu: accurate assembly of segmental duplications, satellites, and allelic variants from high-fidelity long reads. *Genome Res* **30:** 1291–1305. doi:10.1101/gr.263566.120

Nurk S, Koren S, Rhie A, Rautiainen M, Bzikadze AV, Mikheenko A, Vollger MR, Altemose N, Uralsky L, Gershman A, et al. 2022. The complete sequence of a human genome. *Science* **376:** 44–53. doi:10.1126/science.abj6987

Ohtsubo Y, Ikeda-Ohtsubo W, Nagata Y, Tsuda M. 2008. GenomeMatcher: a graphical user interface for DNA sequence comparison. *BMC Bioinformatics* **9:** 376. doi:10.1186/1471-2105-9-376

Olexiouk V, Van Criekinge W, Menschaert G. 2018. An update on sORFs.org: a repository of small ORFs identified by ribosome profiling. *Nucleic Acids Res* **46:** D497–D502. doi:10.1093/nar/gkx1130

Paysan-Lafosse T, Blum M, Chuguransky S, Grego T, Pinto BL, Salazar GA, Bileschi ML, Bork P, Bridge A, Colwell L, et al. 2023. InterPro in 2022. *Nucleic Acids Res* **51:** D418–D427. doi:10.1093/nar/gkac993

Pelletier JF, Sun L, Wise KS, Assad-Garcia N, Karas BJ, Deerinck TJ, Ellisman MH, Mershin A, Gershenfeld N, Chuang R-Y, et al. 2021. Genetic requirements for cell division in a genomically minimal cell. *Cell* **184:** 2430–2440.e16. doi:10.1016/j.cell.2021.03.008

Perez-Riverol Y, Bai J, Bandla C, García-Seisdedos D, Hewapathirana S, Kamatchinathan S, Kundu DJ, Prakash A, Frericks-Zipper A, Eisenacher M, et al. 2022. The PRIDE database resources in 2022: a hub for mass spectrometry-based proteomics evidences. *Nucleic Acids Res* **50:** D543–D552. doi:10.1093/nar/gkab1038

Pertea G, Pertea M. 2020. GFF utilities: GffRead and GffCompare. *F1000Res* **9:** 304. doi:10.12688/f1000research.23297.1

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

Rautiainen M, Nurk S, Walenz BP, Logsdon GA, Porubsky D, Rhie A, Eichler EE, Phillippy AM, Koren S. 2023. Telomere-to-telomere assembly of diploid chromosomes with Verkko. *Nat Biotechnol* **41:** 1474–1482. doi:10.1038/s41587-023-01662-6

Roach NP, Sadowski N, Alessi AF, Timp W, Taylor J, Kim JK. 2020. The full-length transcriptome of *C. elegans* using direct RNA sequencing. *Genome Res* **30:** 299–312. doi:10.1101/gr.251314.119

Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP. 2011. Integrative genomics viewer. *Nat Biotechnol* **29:** 24–26. doi:10.1038/nbt.1754

Shafin K, Pesout T, Lorig-Roach R, Haukness M, Olsen HE, Bosworth C, Armstrong J, Tigyi K, Maurer N, Koren S, et al. 2020. Nanopore sequencing and the Shasta toolkit enable efficient de novo assembly of eleven human genomes. *Nat Biotechnol* **38:** 1044–1053. doi:10.1038/s41587-020-0503-6

Shumate A, Salzberg SL. 2021. Liftoff: accurate mapping of gene annotations. *Bioinformatics* **37:** 1639–1643. doi:10.1093/bioinformatics/btaa1016

Stanke M, Diekhans M, Baertsch R, Haussler D. 2008. Using native and syntenically mapped cDNA alignments to improve *de novo* gene finding. *Bioinformatics* **24:** 637–644. doi:10.1093/bioinformatics/btn013

Sterken MG, Snoek LB, Kammenga JE, Andersen EC. 2015. The laboratory domestication of *Caenorhabditis elegans*. *Trends Genet* **31:** 224–231. doi:10.1016/j.tig.2015.02.009

Sternberg PW, Van Auken K, Wang Q, Wright A, Yook K, Zarowiecki M, Arnaboldi V, Becerra A, Brown S, Cain S, et al. 2024. WormBase 2024: status and transitioning to alliance infrastructure. *Genetics* **227:** iyae050. doi:10.1093/genetics/iyae050

Stewart MK, Clark NL, Merrihew G, Galloway EM, Thomas JH. 2005. High genetic diversity in the chemoreceptor superfamily of *Caenorhabditis elegans*. *Genetics* **169:** 1985–1996. doi:10.1534/genetics.104.035329

Subirana JA, Messeguer X. 2021. DNA satellites are transcribed as part of the non-coding genome in eukaryotes and bacteria. *Genes (Basel)* **12:** 1651. doi:10.3390/genes12111651

Tian R, Rehm FBH, Czernecki D, Gu Y, Zürcher JF, Liu KC, Chin JW. 2024. Establishing a synthetic orthogonal replication system enables accelerated evolution in *E. coli*. *Science* **383:** 421–426. doi:10.1126/science.adk1281

Tuncel A, Pan C, Sprink T, Wilhelm R, Barrangou R, Li L, Shih PM, Varshney RK, Tripathi L, Van Eck J, et al. 2023. Genome-edited foods. *Nat Rev Bioeng* **1:** 799–816. doi:10.1038/s44222-023-00115-8

Tyson JR, O'Neil NJ, Jain M, Olsen HE, Hieter P, Snutch TP. 2018. MinION-based long-read sequencing and assembly extends the *Caenorhabditis elegans* reference genome. *Genome Res* **28:** 266–274. doi:10.1101/gr.221184.117

Venter JC, Glass JI, Hutchison CA III, Vashee S. 2022. Synthetic chromosomes, genomes, viruses, and cells. *Cell* **185:** 2708–2724. doi:10.1016/j.cell.2022.06.046

Vergara IA, Mah AK, Huang JC, Tarailo-Graovac M, Johnsen RC, Baillie DL, Chen N. 2009. Polymorphic segmental duplication in the nematode *Caenorhabditis elegans*. *BMC Genomics* **10:** 329. doi:10.1186/1471-2164-10-329

Wahba L, Hansen L, Fire AZ. 2021. An essential role for the piRNA pathway in regulating the ribosomal RNA pool in *C. elegans*. *Dev Cell* **56:** 2295–2312.e6. doi:10.1016/j.devcel.2021.07.014

Wang C, Wang B, Pandey T, Long Y, Zhang J, Oh F, Sima J, Guo R, Liu Y, Zhang C, et al. 2022. A conserved megaprotein-based molecular bridge critical for lipid trafficking and cold resilience. *Nat Commun* **13:** 6805. doi:10.1038/s41467-022-34450-y

Wei S, Chen H, Dzakah EE, Yu B, Wang X, Fu T, Li J, Liu L, Fang S, Liu W, et al. 2019. Systematic evaluation of *C. elegans* lincRNAs with CRISPR knockout mutants. *Genome Biol* **20:** 7. doi:10.1186/s13059-018-1619-6

Wenger AM, Peluso P, Rowell WJ, Chang PC, Hall RJ, Concepcion GT, Ebler J, Fungtammasan A, Kolesnikov A, Olson ND, et al. 2019. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat Biotechnol* **37:** 1155–1162. doi:10.1038/s41587-019-0217-9

Wootton JC. 1994. Non-globular domains in protein sequences: automated segmentation using complexity measures. *Comput Chem* **18:** 269–285. doi:10.1016/0097-8485(94)85023-2

Xia T, Horton ER, Salcini AE, Pocock R, Cox TR, Erler JT. 2018. Proteomic characterization of *Caenorhabditis elegans* larval development. *Proteomics* **18:** 1700238. doi:10.1002/pmic.201700238

Xie L, Gong X, Yang K, Huang Y, Zhang S, Shen L, Sun Y, Wu D, Ye C, Zhu Q-H, et al. 2024. Technology-enabled great leap in deciphering plant genomes. *Nat Plants* **10:** 551–566. doi:10.1038/s41477-024-01655-6

Yoshimura J, Ichikawa K, Shoura MJ, Artiles KL, Gabdank I, Wahba L, Smith CL, Edgley ML, Rougvie AE, Fire AZ, et al. 2019. Recompleting the *Caenorhabditis elegans* genome. *Genome Res* **29:** 1009–1022. doi:10.1101/gr.244830.118

Yu X, Gray S, Ferreira HC. 2023. POT-3 preferentially binds the terminal DNA-repeat on the telomeric G-overhang. *Nucleic Acids Res* **51:** 610–618. doi:10.1093/nar/gkac1203

Zhao Y, Coelho C, Hughes AL, Lazar-Stefanita L, Yang S, Brooks AN, Walker RSK, Zhang W, Lauer S, Hernandez C, et al. 2023. Debugging and consolidating multiple synthetic chromosomes reveals combinatorial genetic interactions. *Cell* **186:** 5220–5236.e16. doi:10.1016/j.cell.2023.09.025

Received December 5, 2024; accepted in revised form June 6, 2025.