# Recompleting the *Caenorhabditis elegans* genome

Jun Yoshimura,[1,7] Kazuki Ichikawa,[1,7] Massa J. Shoura,[2,7] Karen L. Artiles,[2,7] Idan Gabdank,[3] Lamia Wahba,[2] Cheryl L. Smith,[2,3] Mark L. Edgley,[4] Ann E. Rougvie,[5] Andrew Z. Fire,[2,3] Shinichi Morishita,[1] and Erich M. Schwarz[6]

[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,* [3]*Department of Genetics, Stanford University, Stanford, California 94305, USA;* [4]*Department of Zoology and Michael Smith Laboratories, University of British Columbia, Vancouver V6T 1Z3, British Columbia, Canada;* [5]*Department of Genetics, Cell Biology, and Development, University of Minnesota, Minneapolis, Minnesota 55454, USA;* [6]*Department of Molecular Biology and Genetics, Cornell University, Ithaca, New York 14853, USA*

**_Caenorhabditis elegans_ was the first multicellular eukaryotic genome sequenced to apparent completion. Although this assembly employed a standard _C. elegans_ strain (N2), it used sequence data from several laboratories, with DNA propagated in bacteria and yeast. Thus, the N2 assembly has many differences from any _C. elegans_ available today. To provide a more accurate _C. elegans_ genome, we performed long-read assembly of VC2010, a modern strain derived from N2. Our VC2010 assembly has 99.98% identity to N2 but with an additional 1.8 Mb including tandem repeat expansions and genome duplications. For 116 structural discrepancies between N2 and VC2010, 97 structures matching VC2010 (84%) were also found in two outgroup strains, implying deficiencies in N2. Over 98% of N2 genes encoded unchanged products in VC2010; moreover, we predicted ≥53 new genes in VC2010. The recompleted genome of _C. elegans_ should be a valuable resource for genetics, genomics, and systems biology.**

**[Supplemental material is available for this article.]**

The usefulness of model organisms in modern biology partly comes from their having reference genome assemblies of the highest possible quality. For *Caenorhabditis elegans*, such an assembly for the wild-type strain N2 has existed for 20 yr (The *C. elegans* Sequencing Consortium 1998). However, this assembly was generated with sequence data from N2 and CB1392 [*nuc-1*(*e1392*)] populations of uncertain lineage grown in at least two different laboratories during the 1980s and 1990s (Coulson et al. 1988, 1991, 1995; R Waterston, pers. comm.). Although these populations were originally derived from a single *C. elegans* hermaphrodite selected by Brenner in the 1960s, accuracy of the resulting reference genome is limited both by genetic variants (arising between laboratory N2 strains over decades) (Gems and Riddle 2000; Vergara et al. 2009; Sterken et al. 2015) and by technical limitations (arising from clone-based Sanger technology) (Godiska et al. 2010; for review, see Ross et al. 2013). While no systematic genome-wide comparison of modern N2 laboratory strains has yet been published, N2 may have accumulated up to 1000 neutral mutations even before it was first frozen in 1969 (Sterken et al. 2015), and substantial genetic differences have subsequently arisen between N2 strains in different laboratories (Gems and Riddle 2000; Vergara et al. 2009). Thus, the current *C. elegans* reference genome does not exactly correspond with any N2 strain that exists today. Many genetics, genomics, and systems biology experiments with *C. elegans* would be aided by having a truly isogenic reference strain with a maximally accurate genome assembly.

The N2 reference assembly is putatively gap-free, and ideally its replacement should also be. However, generating gap-free animal genomes is quite difficult. Draft genomes based on Illumina short-read sequencing have highly accurate DNA sequences through the vast majority of the genome but are plagued with gaps that confound gene prediction and analysis (Denton et al. 2014). To fill such gaps, recent studies have assembled long reads of ≥10 kb with programs such as PBcR (Koren et al. 2012), HGAP (Chin et al. 2013), DALIGN (Myers 2014), MHAP (Berlin et al. 2015), FALCON (Chin et al. 2016), miniasm (Li 2016), Canu (Koren et al. 2017), HINGE (Kamath et al. 2017), and MARVEL (Nowoshilow et al. 2018). These programs have been used on Pacific Biosciences (PacBio) reads to provide assemblies for organisms such as *Escherichia coli* (Chin et al. 2013), humans (Berlin et al. 2015; Chaisson et al. 2015; Pendleton et al. 2015), other primates (Gordon et al. 2016; Kronenberg et al. 2018), and *Drosophila* (Chakraborty et al. 2018; Solares et al. 2018). Gap-free bacterial genomes were obtained (Chin et al. 2013), and each assembled vertebrate assembly had extremely long contigs with hundreds fewer gaps than Sanger-based assemblies. Despite considerable progress, it is still challenging to produce gap-free genomes for eukaryotes with genome sizes of 100 Mb or more, in part because of their often long and highly repetitive centromeric regions (VanBuren et al. 2015; Ichikawa et al. 2017). Only one human centromere has been assembled with BAC clones and long-read data so far (Jain et al. 2018b).

*C. elegans* chromosomes are holocentric and lack the highly repetitive centromere regions seen in other animal models (Friedman and Freitag 2017). However, other types of genomic repeats can also prevent gap-free assemblies; for example, a recent long-read-based assembly of *C. elegans* had 48 contigs, with 42 remaining gaps (Tyson et al. 2018). Thus, fully replacing the gap-free

**[7]These authors contributed equally to this work.**
**Corresponding authors: afire@stanford.edu, moris@edu.k.u-tokyo.ac.jp, ems394@cornell.edu**

Article published online before print. Article, supplemental material, and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.244830.118. Freely available online through the *Genome Research* Open Access option.

N2 sequence will require exceptional efforts even after performing long-read genome assembly.

Although the N2 genome assembly has long been considered complete (The *C. elegans* Sequencing Consortium 1998; Hillier et al. 2005), new evidence shows that it still has missing sequences. An early hint of such sequences was the observation of size discrepancies in DNA fragments between Southern blots (250 and 70 kb) versus assembly predictions (6 and 20 kb, respectively) (Hillier et al. 2005). More recently, reassembly of N2 with Illumina synthetic long reads showed at least ~40 kb to be missing from the assembly (Li et al. 2015). Another study used Oxford Nanopore Technology (Nanopore) long-read sequencing (Loman and Watson 2015; Deamer et al. 2016) to produce an assembly with ~2 Mb of additional sequences missing from the N2 reference assembly (Tyson et al. 2018). However, the nucleotide identity of this new assembly with N2 was 99.84% even after polishing with Illumina short reads, with a mismatch ratio of ~0.16% and 42 gaps, making it a less than ideal basis for an updated reference genome. Similar problems were encountered for the Nanopore-derived genome assemblies of *E. coli* (Loman et al. 2015) and *Saccharomyces cerevisiae* (Goodwin et al. 2015), with their respective nucleotide identities to high-accuracy reference genomes at 99.5% and 99.88%. In contrast, the single-molecule real-time long-read sequencing of PacBio (Levene et al. 2003; Korlach et al. 2008; Eid et al. 2009) appears free from sequence bias in its errors (Myers 2014), allowing reliable self-correction; accordingly, PacBio has been used to generate an *E. coli* assembly with an accuracy of 99.9995% (Chin et al. 2013). Given the degree of completeness and accuracy in genome assembly needed for a model system with the array of tools available for *C. elegans*, using all three sequencing technologies (Illumina, PacBio, and Nanopore) would be optimal.

## Results

### Genome sequencing and *C. elegans* strain

We sought to generate a new *C. elegans* assembly that matches a modern and easily available reference strain, VC2010 (Flibotte et al. 2010), a nonmutagenized derivative of N2. We began by sequencing VC2010 with short-read Illumina (genome coverage, 50-fold; mean read length 73 nt), long-read PacBio RSII (genome coverage, 290-fold; mean length, 8.8 kb), and long-read Oxford Nanopore MinION (genome coverage, 32-fold; mean length, 14.2 kb) (Supplemental Table S1).

To ensure reproducibility of this assembly in vivo, we derived a highly clonal strain from VC2010, called PD1074, and used it to generate most of our genomic sequence data. PD1074 has been deposited at the *Caenorhabditis* Genetics Center stock center (CGC; https://cgc.umn.edu/strain/PD1074) and is the recommended strain for biological work using the VC2010 genome assembly described here. *C. elegans* researchers who wish to have significantly higher genomic and genetic reproducibility than is possible with N2 are encouraged to adopt PD1074 as a new reference strain for wild-type controls, classical mutagenesis, and genome engineering.

### Genome assembly and local reassembly

We assembled raw PacBio reads with the long-read genome assemblers Canu (Koren et al. 2017), FALCON (Chin et al. 2016), miniasm (Li 2016), and HINGE (Kamath et al. 2017); each used different algorithms to handle repeats and hence could yield complementary assembly gaps from the same input sequencing data. For genome assembly, we used either PacBio reads alone (which had a lower error rate) or both PacBio and Nanopore reads (which had longer reads). In total, we generated seven assemblies (Supplemental Table S2; https://osf.io/jx89y).

Our genome assembly exhibited some large-scale differences from the N2 reference assembly. These might reflect structural variants (SVs) in vivo or might instead reflect errors of either assembly in silico. To evaluate these possibilities, we inspected possible SVs by aligning raw reads to them and checking for consistency between SVs and aligned reads. In our VC2010 assembly, 136 of 156 SVs were supported by multiple long reads, leaving 20 other SVs that we interpreted as assembly errors. To correct these errors, we collected raw reads that matched each local region and reassembled the reads into contigs with Canu (Supplemental Fig. S1). We found that 14 locally reassembled contigs perfectly matched the N2 reference assembly, while the remaining six contigs had different tandem repeat expansions. Therefore, although de novo genome assembly from long reads may contain false-positive SVs, many of them can be corrected by local reassembly.

### Closing gaps with multiple assemblies

Each VC2010 assembly had 76–202 gaps (median, 111 gaps) (Supplemental Table S2). To fill them, we ordered all contigs of the seven assemblies along the six *C. elegans* nuclear chromosomes and mitochondrial chromosome by aligning them to the N2 reference assembly (Figs. 1A, 2A; Supplemental Fig. S2; Kurtz et al. 2004). Among the seven assemblies, we selected the Canu assembly as primary because it was the largest of the three assemblies using PacBio reads (Supplemental Table S2). We then closed gaps in the Canu assembly with contigs in the remaining six genome assemblies (Fig. 1B,C; Supplemental Fig. S3). A single long contig sometimes could span more than one gap (Fig. 1D). When we could not span a gap with a single contig, we searched for multiple overlapping contigs to fill the gap (Fig. 1E). If that failed, we also used error-corrected reads that we obtained as a byproduct of Canu assembly, by aligning PacBio's raw reads to each other and by taking the consensus sequence in multiple alignments. We aligned error-corrected reads to contigs around unsettled gaps with BLASR (Chaisson and Tesler 2012) and checked if gaps were spanned by multiple error-corrected reads (Fig. 1F). Finally, we tried a hybrid approach of using both contigs and corrected reads to fill gaps (Fig. 1G; Supplemental Fig. S2; Supplemental Table S3).

### Filling five large gaps with Nanopore long reads

This strategy yielded a genome assembly with five large gaps surrounded by long tandem repeat expansions (Supplemental Table S4). We confirmed that pairs of two contigs around the five gaps were in proximity by quantifying Hi-C reads linking these paired contigs (Supplemental Methods). To fill those gaps, we collected extremely long Nanopore reads so that the genome was covered 11-fold by reads with lengths of ≥35 kb (Fig. 2A; Supplemental Table S1). By aligning these Nanopore reads, we could close three gaps. For instance, one 93-kb Nanopore read spanned a gap in Chromosome I (nt 4,605,602–4,668,785) (Fig. 2B,C) and linked two surrounding contigs (Fig. 2D; Supplemental Fig. S4A). Similarly, a gap in Chromosome II (nt 14,525,986–14,566,400) was spanned by two independent Nanopore reads (Supplemental Fig. S4B), and one of the two gaps in Chromosome X (nt 4,149,383–4,226,837) was spanned by two overlapping reads (Supplemental Fig. S4C).

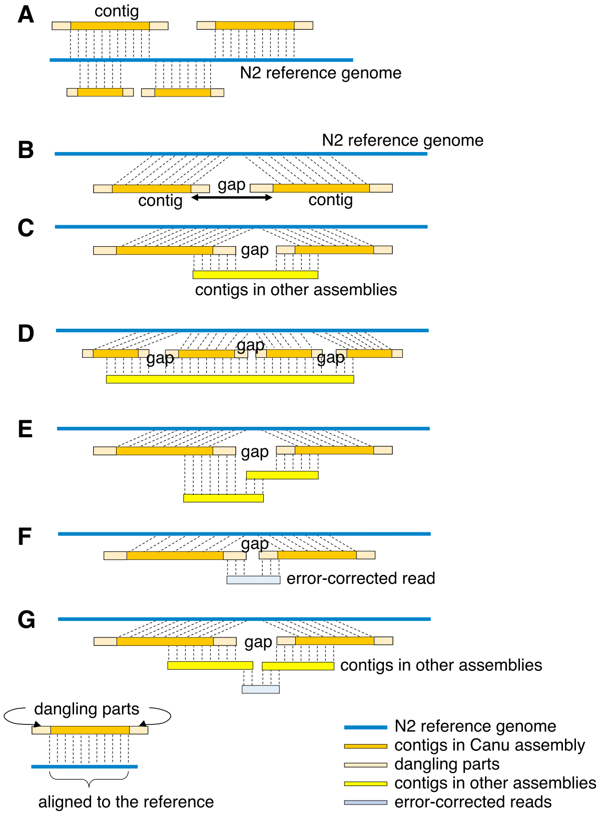

**Figure 1.** Steps for detecting and filling gaps. (*A*) Contigs are ordered along the N2 reference assembly. Parts shown as dangling (colored light orange) fail to align and are missing in the N2 reference. (*B*) At a gap, regions in two Canu contigs (orange) map to proximal loci on the N2 reference; however, the two contigs have dangling end subsequences missing in the reference. In such cases, we estimate gaps between the contigs according to steps illustrated in *C*–*G*. (*C*) A single contig in other assemblies (yellow) fills a gap. (*D*) A long contig in other assemblies combines multiple contigs separated by more than one gap. (*E*) More than one contig fills a gap. (*F*) A single error-corrected read (light blue) fills a gap. (*G*) A hybrid approach of using multiple contigs and error-corrected reads fills a gap.

The remaining two gaps, however, were difficult to close completely. The structure of a gap on Chromosome X (nt 5,215,995–5,284,061) was partly determined (Supplemental Fig. S4D). Another gap in Chromosome III (nt 7,587,315–7,682,900) could not be fully resolved but was estimated to have two tandem repeats of size >40 kb around the gap (Supplemental Fig. S4E). We also attempted to use publicly available Nanopore reads from strain VC2010 (Tyson et al. 2018), but these were shorter than ours and were not effective in closing gaps. Conversely, examining our Nanopore reads that mapped around the final gaps in our genome assembly identified long tandem repeats (Supplemental Table S4) that were underestimated or erroneous in the N2 reference assembly (Supplemental Fig. S5).

## Statistics and accuracy of the final VC2010 assembly

Our final VC2010 genome assembly is summarized in Table 1 (with more extensive data in Supplemental Tables S5 and S6 and Supplemental Data File S1). Its N50 contig length was 15.5 Mb; Chromosomes I, II, IV, and V had no gaps. The nucleotide difference between the N2 reference and VC2010 assemblies was 0.02192%. When we error-corrected nucleotides by aligning Illumina short reads to the assembly using Pilon (Walker et al. 2014), we observed a decrease in the indel ratio by 0.00050% but an increase in the mismatch ratio by 0.00027%, implying that the quality of PacBio assembly alone was very high (Supplemental Table S6). Of 982 single-copy genes used by BUSCO to assess the completeness of genome assemblies (Waterhouse et al. 2018), our VC2010 assembly included 975 complete genes (99.3%) and four fragmented genes; identical frequencies were seen in N2.

As an independent test of our VC2010 assembly's quality, we aligned Illumina reads from two *C. elegans* strains (VC2010 and CB4856, a wild isolate from Hawaii) to the assemblies of N2, Nanopore-based VC2010 (Tyson et al. 2018), and our VC2010; we then determined which assembly exhibited the highest rate of well-mapped reads. The Illumina reads for VC2010, which were collected from the same strain (PD1074) used for the VC2010 reference, had not been previously used to polish either VC2010 assembly and thus were independent of both. For both read sets, our VC2010 assembly exhibited more mapped reads ($P < 10^{-15}$) (Supplemental Methods), fewer clipped nucleotides at 5′ and 3′ ends of reads that failed to map, and more aligned bases compared to N2 or the Nanopore-based VC2010 assembly by Tyson et al. (Supplemental Table S7). The number of aligned bases is an imperfect measure of quality, since mismatched bases include errors in MiSeq reads; however, the other two indicators should be robust to read errors. Out of 6,174,360 Illumina reads, all but 201,926 (3.3%) could be mapped to our VC2010 assembly. Of these 201,926 unmapped reads, 166,407 (82.4%) could be aligned to *E. coli* genomes (Supplemental Methods); since *C. elegans* is routinely grown on *E. coli* as a food source, this result is unsurprising and shows that very few MiSeq reads were truly unmappable.

For the VC2010 assembly to be an improved reference genome assembly, it must retain gene structures and other annotations of the N2 genome into which decades of work have been invested (Spieth et al. 2014; Lee et al. 2018). We used chain alignments (Kent et al. 2003) to lift over gene annotations from N2 to VC2010 (Supplemental Data Files S2, S3) and then determined how many genes still encoded unchanged products (Supplemental Data Files S4, S5; Supplemental Table S8). Out of 20,104 protein-coding genes in WormBase release WS264, all but 214 still encoded at least one unchanged isoform after liftover to the VC2010 assembly. Similarly, out of 25,042 ncRNA-coding genes, all but 530 encoded at least one unchanged isoform in the VC2010 assembly. This left less than 2% of N2 genes that will require manual annotation to be assigned valid structures in VC2010. Unmapped or structurally altered N2 genes arise from sequence differences that can also contain novel genes, as discussed below.

## New genome regions in *C. elegans*

The VC2010 assembly is 1.8 Mb longer than the N2 reference assembly. Although we suspect that these extra sequences may have preexisted in many or all N2-derived strains, we describe them here as new genome regions. These were classified as tandem repeat expansions, insertions, duplications, telomeres, and other smaller differences (Fig. 3A). The most abundant category consisted of 119 large tandem repeat expansions of ≥1 kb that each had nearly identical units in N2 but different numbers of those units; this category accounted for 847 kb of the difference in assembly sizes. These included 10 tandem repeat expansions around the five large assembly gaps (Supplemental Table S4); the remaining 109 are detailed in Supplemental Table S9 and Supplemental Figure S6. Similar repeat expansions were recently observed in a long-read

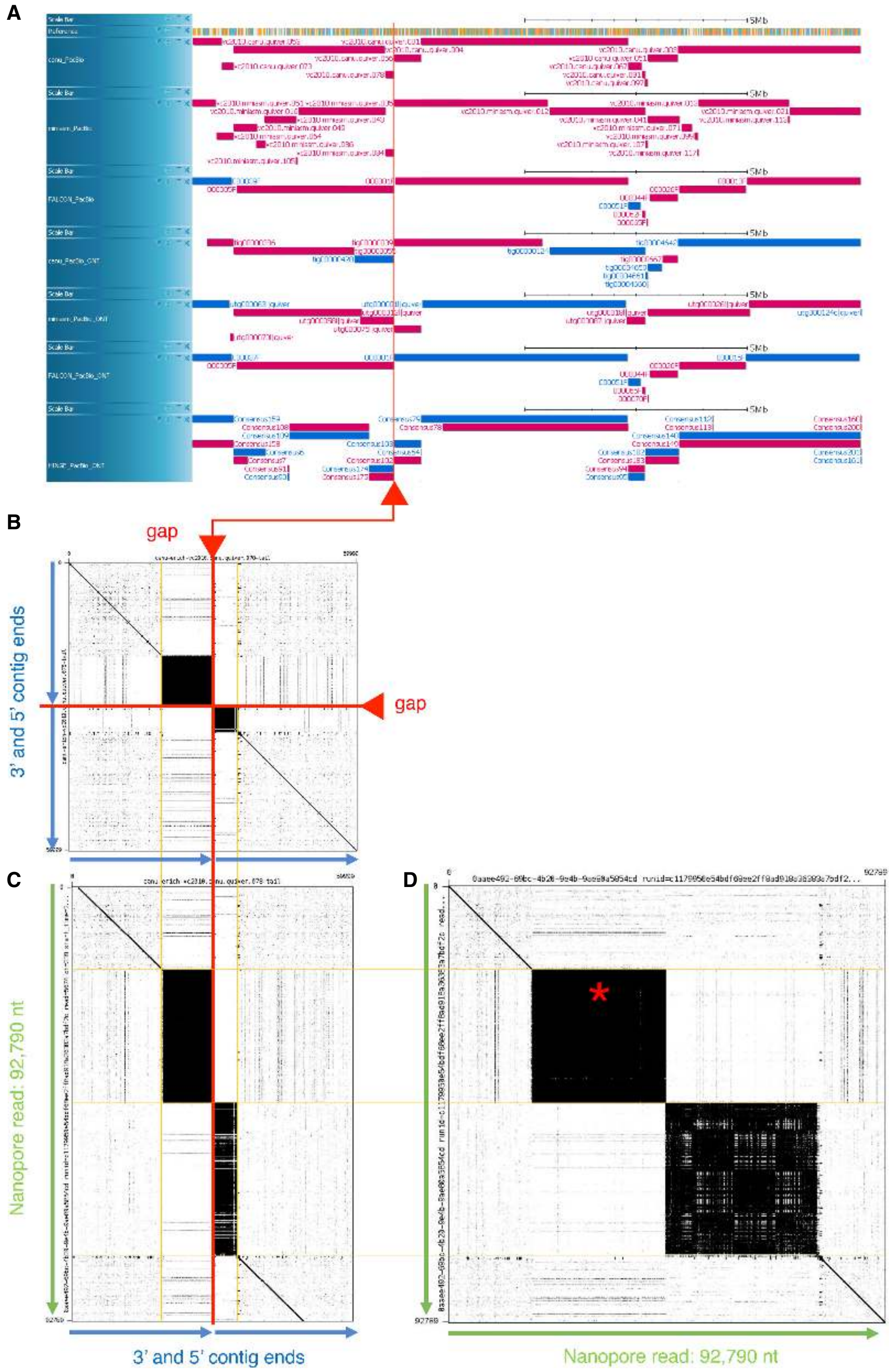

**Figure 2.** Large gaps closed by long Nanopore reads. (*A*) Contigs of seven genome assemblies are aligned with Chromosome I of the N2 reference (see layouts for all chromosomes in Supplemental Fig. S2). The respective red and blue thick lines show alignments of contigs in the plus and minus strands. The vertical red line shows a large gap that failed to be filled by seven genome assemblies. (*B*–*D*) Examples of provisional gap closure using Nanopore data for a region where a long gap was found. (*B*) A self-dot plot for an initial model in which we ligate the last 30 kb of sequence from a contig just before a gap on Chromosome I (colored red) to 30 kb of sequence from another contig just after that gap. Two black boxes represent long tandem repeat expansions around the gap. (*C*) A dot plot between a single 92,790-nt Nanopore read (green) that connects the gap and the simple ligation model in *B*. (*D*) A self-dot plot of the Nanopore read shows that the two tandem repeats in *C* were underestimated. In this example, the left tandem repeat (red asterisk) has 1130 copies of a 26-nt unit string (5′-CATTTTTCTAAAATCCGCCGCAATGC-3′). Supplemental Table S4 shows the units of all tandem repeats in five large assembly gaps.

**Table 1.** **Comparison of *C. elegans* genome assemblies**

| Assembly | N2 (WS264) | VC2010 (Tyson) | VC2010 (ours) |
|---|---|---|---|
| Total size (nt) | 100,286,401 | 103,126,775 | 102,092,263 |
| Scaffolds | 7 | 48 | 7 |
| Contigs | 7 | 48 | 9 |
| Gaps | 0 | 42 | 2 |
| N50 contig size (nt) | 17,493,829 | 4,109,193 | 15,525,148 |
| Max. contig size (nt) | 20,924,180 | 8,173,237 | 21,243,235 |
| Min. contig size (nt) | 13,794 | 42,422 | 13,988 |
| BUSCO complete | 975/982 (99.3%) | 973/982 (99.1%) | 975/982 (99.3%) |
| BUSCO fragmented | 979/982 (99.7%) | 978/982 (99.6%) | 979/982 (99.7%) |
| N2 identity (%) | 100.00 | 99.84 | 99.98 |

Statistics for the N2 reference assembly (WormBase WS264) (Lee et al. 2018), a Nanopore-based VC2010 assembly (Tyson et al. 2018), and our VC2010 assembly are compared. BUSCO coverage is from the BUSCO nematode reference gene set (upper row, complete alignments; lower row, complete or fragmented ones). Identity to N2 is by nucleotides. BUSCO reported that alignments of seven genes were fragmented or missing, but we found complete alignments for the seven genes (Supplemental Methods). For N2 and our VC2010, scaffolds correspond to chromosomes (six nuclear chromosomes and one mitochondrial chromosome). For the VC2010 (Tyson) assembly, identity to N2 was computed previously (Tyson et al. 2018).

assembly for the parasitic nematode *Nippostrongylus brasiliensis* (Eccles et al. 2018) and were classified as very long stretches of complex tandem repeats (VeCTRs). Comparing the regions of tandem repeat expansions in VC2010 to regions of the N2 genome that had been assembled from yeast artificial chromosomes (YACs), we observed that these expansions overlapped YACs with a 1.8-fold higher frequency (40.7%) than expected by chance (22.2%; $P < 2.2 \times 10^{-16}$) (Supplemental Table S10; Supplemental Methods); this suggests that YACs used for N2 sequencing were disproportionately likely to lose tandem repeats through recombination in yeast. We also noticed many tandem repeats that had no clear repeat units and have many rearrangements (see the dot plots between these regions in the N2 and VC2010 reference genomes in Supplemental Fig. S7); these are categorized as imperfect tandem repeats in Figure 3A.

One class of tandem repeats that seem to be substantially underestimated in both the N2 and the VC2010 assemblies are copies of 5S RNA, 18S/28S RNA, and positioning sequence on X (psx1) (Sulston and Brenner 1974; Nelson and Honda 1985; Ellis et al. 1986; Stricklin et al. 2005; Johnson et al. 2006), whose respective unit lengths are 980 nt, 7.2 kb, and 172 nt (Supplemental Table S11). While 26, two, and 106 copies of these repeats were respectively found in the VC2010 assembly with an average similarity of 98.8% by homology search (Supplemental Table S12), 67, six, and 185 copies were observed in single long Nanopore reads with an average similarity of 88.5%, though the numbers of copies were largely inaccurate and the repeat copies were not consecutive but had some long gaps among them (Supplemental Table S13). We therefore searched raw Nanopore and PacBio reads for copies of the three repeats and predicted that the expected numbers of respective occurrences are 144–145, 28–55, and 167, implying that many copies were omitted even from the VC2010 assembly (Supplemental Table S11); however, we could not determine the exact number of repeat copies in each relevant locus.

Telomeric tandem repeats of unit GGCTTA (Wicky et al. 1996; Kim et al. 2003) are much longer in the VC2010 assembly than in the N2 assembly (Supplemental Table S14; Supplemental Fig. S8) in all chromosomes except for the right ends of Chromosomes I and III. Both of the two exceptional right ends terminated with large segmental duplications of >5-kb units in the VC2010 assembly, which presumably hindered genome assemblers from extending the right ends toward telomeric GGCTTA tandem repeats. Segmental duplications proximal to telomeric tandem repeats are known as subtelomere repeat elements in human genomes, and telomeric repeat-containing RNA transcribed from such a telomeric region is involved in telomere regulation (Azzalin et al. 2007; Schoeftner and Blasco 2008; McCaffrey et al. 2017). We were able to extend segmental duplications at the right ends of Chromosomes I and III to include telomeric GGCTTA tandem repeats using a few Nanopore and PacBio raw reads that matched the right ends (Supplemental Fig. S9). Because raw PacBio and Nanopore reads are error-prone, we did not incorporate them into these regions of the VC2010 assembly; however, our analysis suggests that Chromosomes I and III have subtelomere repeat elements that may promote telomere maintenance in *C. elegans*.

We then examined large differences (candidates of insertions, deletions, and duplications) of ≥100 nt in size between the N2 and VC2010 assemblies. As one approach to checking whether those large differences could be due to errors in the new assembly, we generated PacBio-based assemblies of the outgroup strains PD2182 and PD2183 (Supplemental Table S15), derived from the wild isolates CB4856 (Hawaiian *C. elegans* isolate) and MY2, respectively. Since CB4856 and MY2 phylogenetically diverged from N2 long before the divergence of VC2010 (Fig. 3B; Wicks et al. 2001; Swan et al. 2002; Rockman and Kruglyak 2009), sequence features shared by VC2010, PD2182, and PD2183 are expected to be ancestral; thus, areas that differ in the N2 reference assembly from all three may represent either strain-specific polymorphisms or errors in the N2 assembly. We assessed large differences between the N2 reference assembly and our VC2010 assembly, examining 100 candidate insertions, six deletions, and 30 duplications (≥100 nt in size) (Supplemental Tables S16–S18), including duplicated genes (Fig. 3C) and three occurrences of transposons (Tc1, Tc4, and Tc6) (Bessereau 2006; Supplemental Table S19). We compared regions of these candidate insertions, deletions, and duplications by inspecting dot plots between our VC2010 assembly and each of N2, PD2182, and PD2183 (Supplemental Figures S10–S12). Figure 3, C and D, shows a genome duplication that is found only in the VC2010 assembly, being absent from the other three genomes. Figure 3, E and F, shows cases where VC2010, PD2182, and PD2183 appear identical, with genomic regions missing only in the N2 reference assembly. Figure 3G shows the proportions for each class of large differences. Among the examined discrepant regions, assembly deficiencies or losses in N2 are likely to have accounted for 97, while only 19 were identified as bona fide variants unique to the VC2010 assembly.

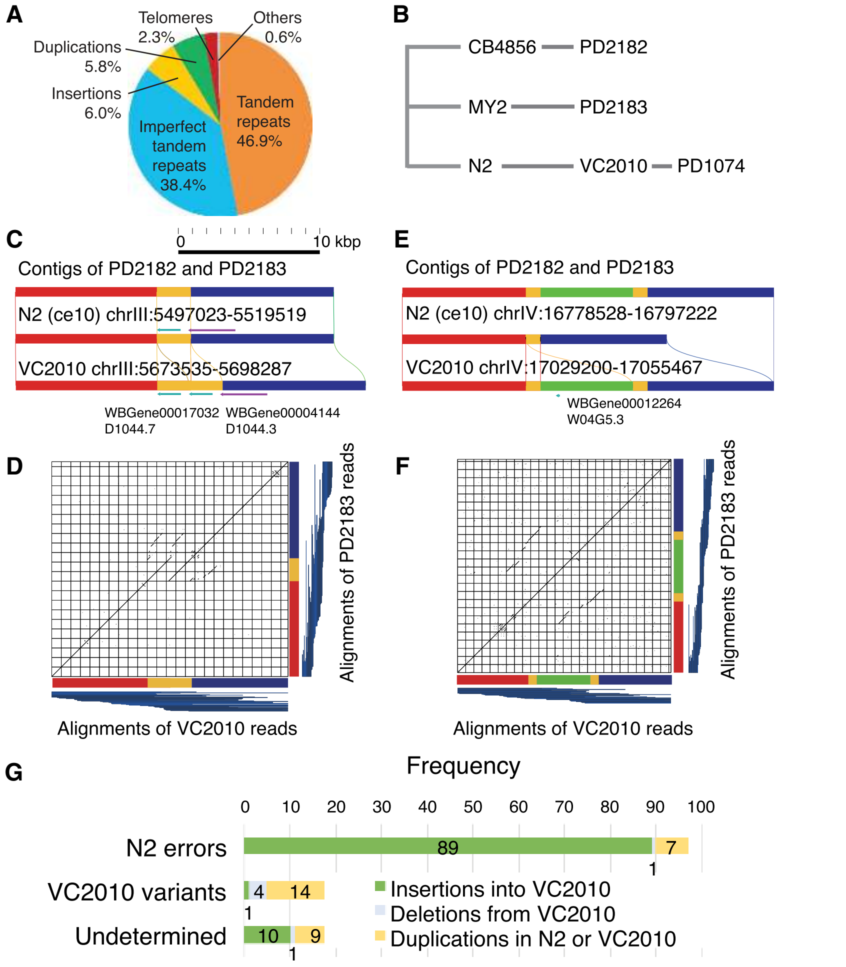

**Figure 3.** New genomic regions in VC2010 assembly. (*A*) Subdivision of sequence classes causing the 1.8-Mb increase in genome size from N2 assembly to VC2010. Large tandem repeat expansions (of size >1 kb) are predominant, accounting for 85% of the increased VC2010 DNA. Other sequence classes include insertions (>100 nt), duplications (>100 nt), and telomere repeats. Tandem repeats are divided into some with clear repeat units and others ("imperfect") without them (Supplemental Fig. S7). (*B*) Phylogenetic tree of N2, VC2010 (PD1074), and outgroup strains CB4856 (PD2182) and MY2 (PD2183). (*C*) The yellow-colored duplicated region with two copies of a gene in VC2010 is compared with its best matching regions in N2, PD2182, and PD2183. The comparison implies that the duplication was a recent event occurring in the lineage from the original N2 strain to VC2010. Of note, two duplicated regions overlap slightly. (*D*) Because long reads were unavailable for N2, we compare the regions in VC2010 and PD2183 for which long reads were available, and we show a dot plot between the regions (a similar dot plot between VC2010 and PD2182 is shown in Supplemental Fig. S12). To confirm the correctness of both regions, we align raw PacBio reads collected from VC2010 and PD2183 to their respective genomic regions, and the alignments are shown as blue lines *below* the *x*-axis and to the *right* of the *y*-axis. Indeed, a number of alignments span and validate the focal duplicated region and its matching region. (*E*) A comparison of regions where VC2010, PD2182, and PD2183 coincide, but the green-colored region is missing in the N2 reference assembly, implying that the segment had been lost in culturing animals or clones used for the N2 assembly or in the original N2 assembly process. (*F*) As in *D*, aligning raw PacBio reads to both regions in VC2010 and PD2183 shows their validity (a similar dot plot between VC2010 and PD2182 is shown in Supplemental Fig. S10). (*G*) Frequencies of apparent insertions into VC2010 (missing in N2), deletions from VC2010 (surplus in N2), and genome duplications (in N2 or VC2010), sorted into three categories: 97 assembly errors in the N2 genome, 19 variants that arose in the lineage from N2 to VC2010, and 20 undetermined cases because of inconsistency among the four genomes. We categorized individual large variants by inspecting the dot plots in Supplemental Figures S10–S12 (Supplemental Tables S16–S18). Of the 97 assembly errors, 89 (92%) were regions missing in the N2 reference assembly.

In lifting reference N2 gene annotations onto VC2010, structural variants would certainly have the potential to affect gene structures (Supplemental Figures S10–S12). Figure 3, C and D, shows one example in which VC2010 evidently harbors multiple copies of a gene while the N2 assembly shows only one copy. Out of 19 confirmed structural variants in VC2010, 14 (74%) were duplications that have the potential to encode near-identical paralogs of *C. elegans* genes (Fig. 3G; Supplemental Table S18).

## New genes in *C. elegans*

To find possible new genes in the VC2010 assembly, we predicted 21,070 protein-coding genes in the VC2010 assembly with AUGUSTUS (Supplemental Data File S6; Stanke et al. 2008). Of these predictions, 19,032 (90%) shared at least one in-frame coding nucleotide with previously annotated, lifted-over N2 genes (Supplemental Table S20). This included 65 N2 reference genes whose coding frames had been garbled by liftover, and these may indeed have been corrected by AUGUSTUS. Conversely, 183 genes predicted by AUGUSTUS had no shared coding nucleotides with lifted-over genes, while also having at least one nucleotide of VC2010-assembly-specific coding DNA (Supplemental Table S21). Examining these, we noted that 13 were entirely contained by lifted-over genes on the same DNA strand to which they showed high similarity (Supplemental Table S22); these 13 nominal genes may actually include alternative exons of the gene encoded by VC2010-assembly-specific DNA. For instance, one such possible exon fell within the *C. elegans* gene for titin (*ttn-1*). More generally, AUGUSTUS gene predictions can differ from those in the N2 reference assembly solely in having additional VC2010-specific exons (Fig. 4A). Of the remaining 170 AUGUSTUS predictions, we considered 117 less likely to be new genes either because they had ≥95% identity to N2 reference protein-coding genes (which might be cases where a preexisting gene had not been properly lifted over) or because they encoded proteins with anomalous compositions (≥50% low-complexity residues or >20 predicted transmembrane sequences, either of which might be mispredicted from highly repetitive DNA). It should be noted that some genes with ≥95% identity might be authentic paralogs of N2 reference genes in very recent genome duplications, but demonstrating such paralogy will require careful future study.

Even after these caveats, there remained 53 AUGUSTUS predictions that we considered most likely to be new genes

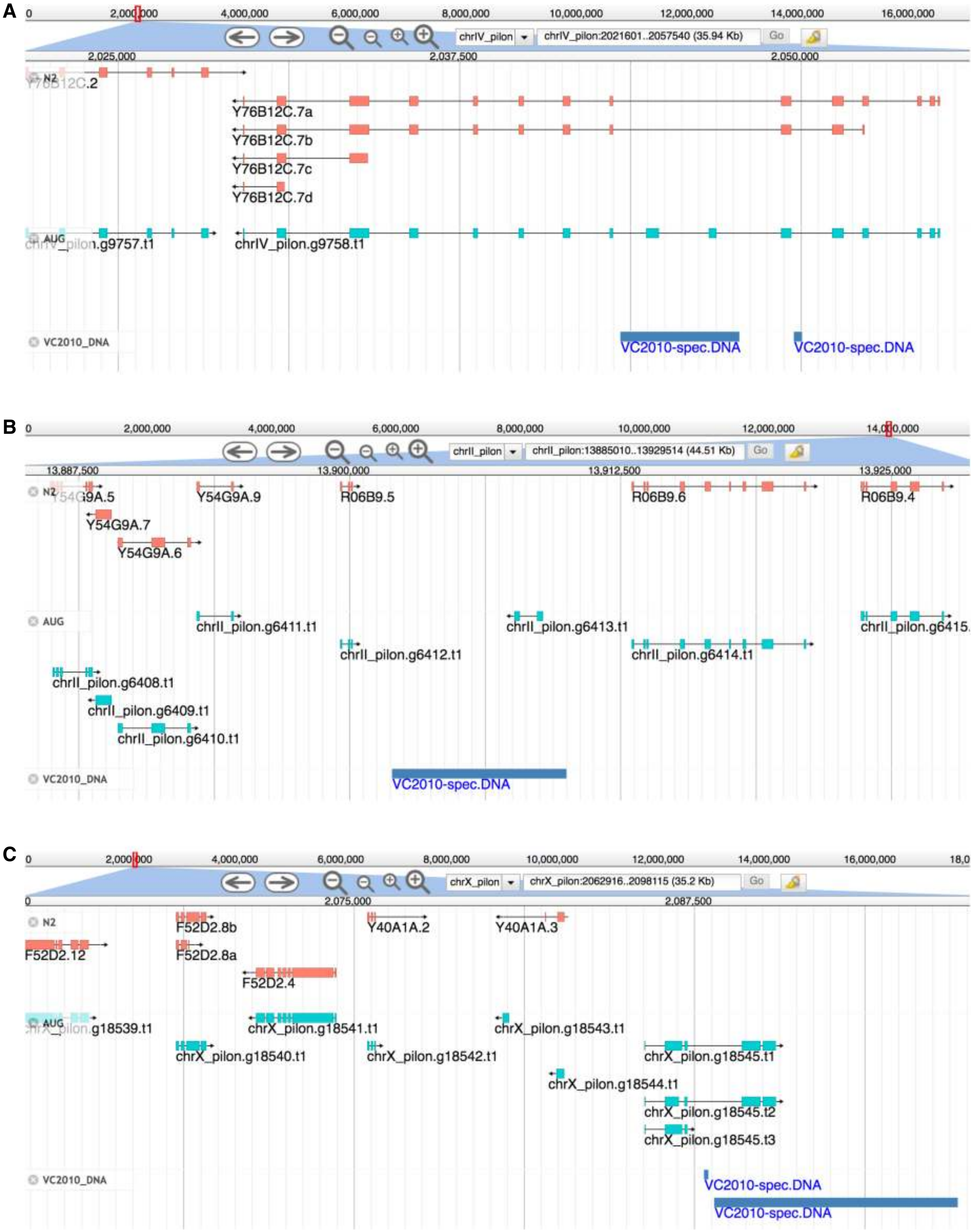

**Figure 4.** New exons and genes in the VC2010 assembly. Segments of the VC2010 assembly are shown with N2-derived gene predictions, independent AUGUSTUS-derived gene predictions, and VC2010-specific DNA regions. For each gene, alternative transcript isoforms (if any) are shown. (*A*) Extra, VC2010-assembly-specific exons in the gene *cpsf-1*/*Y76B12C.7* (alias *chrIV_pilon.g9758*) (Supplemental Table S21). (*B*) *chrII_pilon.g6413*, a likely new gene encoded entirely by VC2010-specific DNA; BLASTP shows this to be a paralog of *T18D3.9*/*MPV17* in the N2 reference assembly but an ortholog of *Cnig_chr_II.g6634* in the PacBio-sequenced *C. nigoni*. Surrounding AUGUSTUS predictions in genomic DNA shared with N2 match N2 reference gene structures closely. (*C*) *chrX_pilon.g18545*, a paralog of *hasp-1*/*C01H6.9* encoded largely by VC2010-assembly-specific DNA. The latter two genes are listed in Supplemental Table S23.

(Supplemental Table S23). One instance of such a gene is *chrII_pilon.g6413*, which is entirely contained within VC2010-specific DNA and encodes a full-length paralog of the mitochondrial inner membrane protein T18D3.9/MPV17 (Fig. 4B); this and other such novel predictions were embedded among AUGUSTUS gene predictions that closely matched N2 reference gene structures (Fig. 4C). We also predicted ncRNA-encoding genomic regions with INFERNAL/RFAM (Supplemental Data File S7), identifying 29 loci that had no overlap with N2 reference liftovers while having at least one VC2010-assembly-specific coding nucleotide (Supplemental Table S24); as noted above, these predominantly included 5S rRNAs and tRNAs, but we also observed loci encoding 5.8S rRNA, large subunit rRNA and small subunit rRNA. Thus, the VC2010 genome assembly will provide a more complete basis than N2 for studies of gene function in *C. elegans*.

## Discussion

The VC2010 assembly presented here, combined with its matching strain PD1074, comprises a tool that will facilitate both genome-wide and individual gene analysis. Although we do not expect this or any assembly to be perfect, the VC2010 assembly provides substantial advantages over its predecessors in both precision and completeness. Reassembly of the *C. elegans* genome was originally intended to produce an improved version of the existing N2 assembly. An actual VC2010 assembly with no gaps was difficult, yielded features not visible in the N2 reference assembly, and has several technical and biological implications.

In recompleting the *C. elegans* genome, we found that different assembly programs could yield sequences with complementary gaps. This proved to be essential for driving the number of gaps down from ~100 to two. In this project, we had the advantage of starting with the N2 reference assembly, which gave us a preexisting standard against which we could compare and validate our multiple VC2010 assemblies. Other genome projects will not always have such a standard. However, merging alternative long-read assemblies also improved genome contiguity in *Drosophila* (Chakraborty et al. 2018; Solares et al. 2018), showing that this may be a generally useful tactic. To close gaps that could not be resolved by comparing PacBio-based assemblies, it proved very valuable to obtain ultralong Nanopore reads. Recent work with Nanopore has given occasional instances of reads up to a megabase long (Jain et al. 2018a), and it seems likely that ultralong reads will become a key tool for completing eukaryotic genomes. Error rates of long reads do matter, however; we were able to do reliable liftover of gene annotations from N2 to VC2010 only because our PacBio reads gave us assemblies with error rates lower than those achievable with Nanopore alone. The optimal mixture of long reads for genome assembly is likely to change rapidly as the error rates of these competing technologies move downward. Another rapidly changing factor will probably be development of new computational tools for sorting and assembling long reads from repeat-rich genomic regions. For example, such repeat-aware assembly methods have recently been used to assemble 10 Mb of heterochromatic sequences from the Y Chromosome of *Drosophila melanogaster* (Chang and Larracuente 2019) and 33–79 Mb of previously unknown paralogous genomic duplications in humans (Vollger et al. 2019). Concerted improvement of long sequencing reads and repeat-aware assembly methods should allow future versions of the *C. elegans* reference genome to be entirely gap-free.

Almost 2% of the putatively gap-free *C. elegans* genome proved to be missing from the N2 assembly; these missing sequences included long stretches of repetitive DNA. Such regions were recently observed in a Nanopore-based assembly of *C. elegans* (Tyson et al. 2018); moreover, they were observed in a Nanopore-based assembly of the parasitic nematode *Nippostrongylus brasiliensis* (Eccles et al. 2018). Given these results and previous analysis of vertebrate genome assemblies (Alkan et al. 2011), it seems likely that most of the nematode assemblies generated over the last decade are missing some repetitive regions of genomic DNA (Korhonen et al. 2016). With the possible exception of highly reduced genomes such as *Pratylenchus coffeae* (Burke et al. 2015), long-read assembly will probably be needed to detect and resolve these systematically lost genome sequences. Long-read genome assembly should also allow the full detection of multigene families residing in tandemly repeated regions of nematode genomes. For parasites such as *N. brasiliensis*, such highly tandem regions may be important for identifying rapidly evolving virulence factors (Raffaele and Kamoun 2012); for free-living nematodes such as *C. elegans*, such regions may be crucial for understanding fast-evolving gene families relevant to nematode ecology (Frezal and Felix 2015).

For the novel repetitive DNA sequences that are noncoding, it is not clear whether there are functions in vivo or not. The case against function is that eukaryotic genomes are shaped not merely by selection but also by mutation and genetic drift that prevent loss of DNA (Lynch 2007). At the same time, eukaryotic cells must suppress the recombination of highly repetitive genome sequences to prevent aneuploidy arising from their unequal crossing over (Charlesworth et al. 1986). Given both of these factors, highly repetitive genomic sequences in nematodes may be nonfunctional but inevitable. On the other hand, many nematode chromosomes are holocentric, lacking classical centromeres to which kinetochores can attach during cell division (Friedman and Freitag 2017); perhaps noncoincidentally, the genomes of holocentric nematodes also have unusually large numbers of repetitive (satellite) DNA elements (Subirana and Messeguer 2013). It is thus possible that repetitive nematode genome sequences serve a biological function, perhaps by providing quasi-centromeric elements on which kinetochores can form.

Finally, recompleting the *C. elegans* genome should make its analysis more effective in several ways. Cloning of mutations by whole-genome sequencing, an increasingly common tool for linking classical genetics to molecular biology (Doitsidou et al. 2016), can now be performed on a truly isogenic reference strain, with mutations being mapped onto a genome that is entirely complete. Analyses of *C. elegans* diversity and population genetics (Cook et al. 2017) can now use a reference genome with tandemly repeated regions and recently evolved genes that are likely genetic hotspots for evolution. Systemic analyses of gene content in *C. elegans* will include a truly full gene and exon complement, and the sequence of genomic DNA for its reference strain will be much less variable between different laboratories; these findings will enable future systems biology of *C. elegans*, such as reengineering of its genome to test possible functions of both its repetitive and nonrepetitive DNA (Richardson et al. 2017).

## Methods

### Genomic DNA extraction

For PacBio DNA sequencing, worm strains were grown on enriched NGM plates prepared with an OP50 bacterial lawn. M9 buffer (22 mM $KH_2PO_4$, 42 mM $Na_2HPO_4$, 86 mM NaCl, 1 mM $MgSO_4$)

(Brenner 1974) was used to wash the animals off the plates before starvation. To ensure minimal bacterial contamination, animals in M9 were then sedimented at ∼450*g* through a cushion of 30% sucrose clinical swinging bucket centrifuge, 50 mL plastic tubes). Genomic DNA was isolated from living and flash-frozen animals approximately as previously described (Shoura et al. 2017). For Nanopore DNA sequencing, we aimed to extract genomic DNA from *C. elegans* with exceptionally low fragmentation and high molecular weight, approximately as previously described (Schwartz and Cantor 1984). For Illumina DNA sequencing, DNA extraction was carried out approximately as previously described (Sha et al. 2010). Further details are given in Supplemental Methods.

### Nanopore library preparation and sequencing of PD1074 (VC3510)

Ultralong single-molecule sequencing of PD1074 genomic DNA was carried out on an Oxford Nanopore MinION sequencer (version Mk1B). A range of Oxford Nanopore library preparation kits and flow cell types were used to accumulate the data summarized in Supplemental Table S1. Among the trials conducted, the highest sequencing yields were obtained using agarose-plug-isolated ultralong DNA prepared with the Oxford Nanopore RAD003 library kit.

Library kit type, flow cell type, and input DNA quantity (as measured by a Qubit fluorometer) are listed for each MinION sequencing run. Run statistics were generated using Poretools (Loman and Quinlan 2014) on FAST5 files which passed filter.

### Selection of N2 and VC2010 genome assemblies for analysis

All of the analyses described here used our final version of the VC2010 genome assembly (version number 20180405; vc2010.draft-20180405.pilon.fasta, provided as Supplemental Data File S1). For large-scale structural comparisons of the N2 reference assembly versus our final VC2010 assembly, we used the version of N2 from WormBase release WS220 (Lee et al. 2018). This has the advantage of being a highly used version that is available in the UCSC Genome Browser (where it is called ‘ce10’). For detailed analyses of exactly which nucleotides in the N2 reference assembly corresponded with nucleotides in VC2010 (particularly, for chain-alignments, lifting over gene annotations, and deciding exactly which nucleotides would be classified as VC2010-assembly-specific), we instead used the most recent version of N2 available at the time of analysis, from WormBase release WS264 (*ftp://ftp.wormbase.org/pub/wormbase/releases/WS264/species/c_elegans/PRJNA13758/c_elegans.PRJNA13758.WS264.genomic.fa.gz*). The most recent change to the N2 genome sequence in WormBase (as of mid-February 2019) was in release WS235; changes of the N2 assembly between WS220 and WS235 (and thus, between WS220 and WS264) were minor. For assessments of Illumina read-mapping frequencies for the different assemblies, we obtained the previously published Nanopore-based VC2010 assembly of Tyson et al. (2018) from the European Nucleotide Archive (ENA) (*ftp://ftp.sra.ebi.ac.uk/vol1/ERA984/ERA984123/oxfordnanopore_native/pilon_4x_polished_assembly_N2_ chips_114_115.tar.gz*); this assembly contains both *C. elegans* and bacterial contigs. For generating the statistics in Table 1, we obtained a version of this assembly with only *C. elegans* contigs from Tyson et al. (2018), archived at https://osf.io/dbgkm.

### Base calling and error correction of PacBio and Nanopore reads

We called bases of PacBio reads using SMRT Analysis 2.3.0/SMRT Pipe 1.87.139483 and, afterward, corrected errors in PacBio reads using the correction and trimming steps of Canu version 1.3 (Koren et al. 2017). Further details are given in Supplemental Methods.

### Assembling PacBio and Nanopore reads using four long-read genome assemblers

For assembling PacBio and Nanopore reads, we used Canu version 1.3 (Koren et al. 2017), miniasm version 0.2 (Li 2016), and HINGE version 0.4.2 (Kamath et al. 2017) with default parameter settings. We also used FALCON version 0.3.0 (Chin et al. 2016) with default parameter settings except for setting the length cutoff for seed reads to 18 kb.

### Aligning genome assemblies

We aligned all contigs in the seven assemblies to the N2 reference assembly (WormBase release WS220; i.e., UCSC ce10) with MUMmer 3.23 (Kurtz et al. 2004), using the program *nucmer* with the arguments *mum mincluster 100 maxgap 300*.

### Assessing effects of Pilon polishing on the VC2010 genome assembly

We used QUAST 4.4 (Gurevich et al. 2013) to calculate mismatch ratios and indel ratios of the VC2010 assembly with respect to the N2 assembly, both before and after polishing of VC2010 with Pilon (Walker et al. 2014) and Illumina short reads.

### Visualizing comparisons of sequence data

For visualizing comparisons of genomic sequences (either from two assemblies or from other sequence data), we generated dot plots using Gepard 1.40 (Krumsiek et al. 2007).

### Analyzing long reads with tandem repeats

Alignment tools such as BLASR are not good at aligning reads with tandem repeats, presumably due to the difficulty in generating a correct chaining in the presence of extensive tandem repeats. We therefore determined the unit string and length of each tandem repeat occurrence, and we aligned reads with tandem repeats to regions with similar tandem repeat patterns around each gap.

### Estimating tandem repeats surrounding five large gaps

We developed a program for calculating the repeat unit string and the number of repeat unit occurrences in each candidate region (https://github.com/morisUtokyo/mTR). We then measured the similarity among repeat unit occurrences in a tandem repeat by using the match ratio of an optimal local alignment between the candidate tandem repeat region and an ideal tandem repeat with no mismatches (a series of tandem repeat unit copies). Because the similarity was higher in PacBio contigs than in raw PacBio/Nanopore reads, we used the repeat unit string calculated from PacBio contigs. To determine the number of repeat unit occurrences, when multiple reads span a tandem repeat, we used the longest tandem repeat. We visualized dot plots of regions from the N2 reference assembly (ce10) and VC2010 assembly around each discordant region together with reliable alignments using Ribbon with default parameters (Nattestad et al. 2016).

### Identifying structural differences of VC2010 from N2 assemblies

To identify structural differences between the N2 and VC2010 genome assemblies, we aligned the genomes with MUMmer version 3.23, input the alignments into Assemblytics (http://assemblytics.com) to call structural variants (apparent genome duplications, insertions into or deletions from the VC2010 genome) between the two genomes, and treated them as candidate different regions. To confirm the presence of each difference, we first corrected raw long reads from VC2010 using the correction and trimming phase in Canu assembler (version 1.4), aligned the corrected long reads to both genomes using BLASR, and used such reliable alignments that their insertions or deletions were of length ≤15 nt, where the upper limit was selected after manual inspection of alignments. Each difference is a pair of reciprocally best matching positions in the VC2010 and N2 assemblies; for such a pair, we expected that the position in the VC2010 assembly would be covered by VC2010 long reads, but the corresponding position in the N2 genome would not be. Indeed, we observed this result in all 107 insertions (Supplemental Fig. S10), all six deletions (Supplemental Fig. S11), and all 30 duplicated regions (Supplemental Fig. S12) except for six large ones (I: 4,400,490–4,420,623, III: 1,296,384–1,341,443, III: 8,571,541–8,624,246, V: 1,5541,324–15,588,233, V: 18,953,527–19,017,461, and V: 19,915,891–19,956,908). We also aligned the WormBase gene set (release WS264) to the VC2010 assembly with minimap2 (splice option) (Li 2018) to examine whether genes were present in the different regions (Supplemental Figs. S10–S12) and whether their structures might be affected by structural variants between VC2010 and N2.

## Searching and estimating repeats of 5S RNA, 18S/28S RNA, and pSX1

We first searched the VC2010 assembly for repeats of 5S RNA (980 nt), 18S/28S RNA (7203 nt), and pSX1 (172 nt), using BLAST with default parameters, and selected such occurrences that the coverage of the query in the alignment was >98% and the identity was >97%. The positions and nucleotide identities of all qualified copies are found in Supplemental Table S12.

We then aligned the three tandem repeats with all raw PacBio/Nanopore reads. Assuming a ∼20% error rate in raw reads, we used BLAST with the default setting of *k*-mer to 11-mer because the sensitivity of detecting matches by BLAST with 11-mer was sufficiently high (≥87.1%) when the query length is ≥100 nt, and an error rate was 20% (Kasahara and Morishita 2006). Although many BLAST alignments were partial, we selected alignments for which ≥90% of the query tandem repeat was aligned and the match ratio was ≥75% given a ∼20% error rate. For respective repeats, we identified the raw reads with the maximum numbers of copies. Supplemental Table S13 shows all copies that met the above conditions.

Lastly, we estimated the expected number of tandem repeat occurrences in the VC2010 worm genome as the number of tandem repeat occurrences in all raw reads divided by the raw read coverage. We calculated the read coverage for 18S/28S RNA carefully because it is 7203 nt long and thus was longer than many raw reads. A raw read that covered one instance of tandem 18S/28S RNA had to be ≥7203 nt in length and was long enough to cover one if its length was 2 × 7203 nt. Thus, to compute the read coverage, we considered two sets of raw reads whose lengths were either ≥7203 nt or ≥14,406 nt for 18S/28S RNA.

Supplemental Table S11 presents the numbers of copies in our assembled VC2010 genome, the maximum numbers of copies in single raw Nanopore reads, and the expected numbers of copies in VC2010.

## Identifying telomeric unit strings

For telomeric tandem repeats at chromosomal ends in the N2 and VC2010 assemblies, we calculated the unit string GCCTAA and the number of unit occurrences at each end using Tandem Repeats Finder (TRF) 4.09 (Benson 1999).

## Lifting over genes from N2 to VC2010

We lifted over gene structures and other genome annotations from the N2 reference assembly to our VC2010 assembly. Such cross-assembly mapping typically requires an annotation file in standard format (e.g., GFF3; https://github.com/The-Sequence-Ontology/Specifications/blob/master/gff3.md), a chain alignment (Kent et al. 2003), and a program capable of mapping annotations based on chain aligments (e.g., liftOver) (Speir et al. 2016). For the N2 genome sequence, we downloaded ftp://ftp.wormbase.org/pub/wormbase/releases/WS264/species/c_elegans/PRJNA13758/c_elegans.PRJNA13758.WS264.genomic.fa.gz. For annotations of canonical N2 genes, we downloaded ftp://ftp.wormbase.org/pub/wormbase/releases/WS264/species/c_elegans/PRJNA13758/c_elegans.PRJNA13758.WS264.canonical_geneset.gtf.gz. Both the genome sequence and its annotations were from the WS264 release of WormBase (Lee et al. 2018). To chain-align our VC2010 assembly (as the query) to the N2 reference assembly (as the target), we used methods almost identical to those described by UCSC for same-species genomic chain alignments (http://genomewiki.ucsc.edu/index.php/Same_species_lift_over_construction). The key differences were as follows: In SameSpeciesBlatSetup.sh, we revised the parameters targetChunkSize and queryChunkSize from 10,000,000 to 22,000,000; we corrected a typographical error from "*B=`basename* [*file*]" to "*B=`basename* [*file*]`"; we used only alignments of homologous chromosomes to build the chain (i.e., only aligned Chromosome I in the VC2010 assembly to Chromosome I in the N2 reference assembly, etc.); and we replaced BLAT alignments (automatically generated in PSL format) with minimap2 alignments (originally generated in SAM format, then reformatted to PSL before chain-building). The liftOver protocol required several utility programs from UCSC; we downloaded some (including liftOver) as precompiled binaries (from http://hgdownload.cse.ucsc.edu/admin/exe/linux.x86_64 and http://hgdownload.cse.ucsc.edu/admin/exe/linux.x86_64/blat), while obtaining others (e.g., endsInLf and partitionSequence.pl) as source code (from http://hgdownload.cse.ucsc.edu/admin/jksrc.v362.zip, http://genomewiki.ucsc.edu/images/9/91/SameSpeciesBlatSetup.sh.txt, and http://genomewiki.ucsc.edu/images/d/d3/SameSpeciesChainNet.sh.txt), and compiling endsInLf from source. We tried chain-alignments based both on BLAT (the original UCSC procedure) and minimap2 (our revised procedure). We noted a slight advantage to the minimap2-based chain alignments in lifting over annotations and so used minimap2 in our final methods. We ran minimap2 (version 2.10-r763-dirty) with the arguments '*-a -t 8 -x asm5*'. We reformatted minimap2 alignments from SAM to PSL with sam2psl.py, downloaded from https://github.com/ndaniel/fusioncatcher/blob/master/bin/sam2psl.py (Nicorici et al. 2014). For mapping genome annotations, we used liftOver with the arguments '*-gff -minMatch=0.90*'.

## Analysis of lifted-over N2 genes

We split canonical N2 genes into three groups for analysis: protein-coding genes with biotype "protein_coding," pseudogenes with biotype "pseudogene," and ncRNA genes having any other biotypes (antisense_RNA, lincRNA, miRNA, ncRNA, piRNA, rRNA, snoRNA, snRNA, or tRNA). To determine the sequences of peptides encoded by protein-coding genes, we extracted their CDS annotations from a GTF annotation file (such as c_elegans.PRJNA13758.WS264.canonical_geneset.gtf, or its equivalent after liftover to VC2010) with awk '($3 == "CDS")' and generated their predicted protein sequences with gffread 0.9.9 (https://github.com/gpertea/gffread). To determine the nucleotide sequences of pseudogenes and ncRNA genes, we extracted their exon annotations (with awk '[$3 == "exon"]') and determined their predicted spliced CDS DNA sequence with gffread. For both extractions, it was important to use CDS or exon records alone in order to avoid segmentation faults with gffread (https://github.com/kingsfordgroup/sailfish/issues/21). The arguments used for protein sequence extraction with gffread were ‘*-C -V -J -M --no-pseudo -g* [*genome assembly*] *-y* [*peptide FASTA output*]’. For extracting the spliced exon sequences for ncRNA genes, we used the arguments ‘*-M --no-pseudo -g* [*genome assembly*] *-x* [*CDS DNA FASTA output*]’; for extracting the spliced exon sequences of pseudogenes, we used almost identical arguments, but with ‘*--no-pseudo*’ omitted. To determine which protein or DNA sequences had changed after liftover for a given gene class (protein-coding, ncRNA, or pseudogenic), we compared pre- and post-liftover sequences with cd-hit-2d from CD-HIT 4.6.8 (Fu et al. 2012), using the arguments ‘*-d 100 -c 1.0 -M 0 -T 1 -l 5 -s 1.0 -aL 1.0 -aS 1.0*’. We extracted gene counts of unchanged and changed sequence products with Perl. Using these methods with both pre-lifted-over N2 gene annotations and their post-lifted-over VC2010 equivalents allowed us to determine how many genes had been lifted over perfectly (because their peptide or DNA sequences were entirely unchanged), how many protein-coding genes had lost a correct coding sequence in the liftover, and which genes remained valid but had some change to their peptide or DNA sequence due to the liftover.

### Delineating new VC2010-assembly-specific genome sequences

Chain alignments are not easily intelligible and cannot be used for comparing genomic locations with programs such as BEDOPS or BEDTools (Quinlan and Hall 2010; Neph et al. 2012). Therefore, we used our minimap2-based chain alignment to build a GFF3 annotation file with exact coordinates for all VC2010 assembly regions either corresponding or not corresponding to blocks in N2. To do this, we first used gen_nt2gff3.pl and the N2 genome sequence (c_elegans.PRJNA13758.WS264.genomic.fa) to build a naive GFF3 ‘annotation’ file for N2 (c_elegans.PRJNA13758.WS264.genomic.gff3) in which every nucleotide of the N2 genome was given a single annotation line, along with a comment stating its original location in N2. We then mapped this one-line-per-nucleotide GFF3 file to our VC2010 assembly with liftOver and our minimap2-based chain alignment, which generated an equivalent one-nt-per-line GFF3 file (c_elegans.PRJNA13758.WS264.genomic.V2010_remap.gff3) now having VC2010 coordinates but with appended annotations for the original N2 coordinates. Finally, we condensed this into a more orthodox and useful GFF3 annotation file (c_elegans.PRJNA13758.WS264.genomic.V2010_remap_blocks.gff) in which every contiguous block in the VC2010 assembly was either annotated as mappable from the N2 reference assembly (with a note showing its original coordinates) or nonmappable from N2; to do this, we used block_summarize_1nt_gff.pl with the arguments ‘*-p nucleotide_match -n region -d vc2010.draft-20180405.pilon.fasta -g c_elegans.PRJNA13758.WS264.genomic.V2010_remap.gff3*’, then revising the original output with ‘*cat c_elegans.PRJNA13758.WS264.genomic.V2010_remap_blocks.orig.gff | revise_liftover_blocks_10jun2018.pl > c_elegans.PRJNA13758.WS264.genomic.V2010_remap_blocks.gff*’. The final block annotation file was later used to identify genes that fell entirely within novel, VC2010-assembly-specific genomic regions. The Perl scripts gen_nt2gff3.pl, block_summarize_1nt_gff.pl and revise_liftover_blocks_10jun2018.pl are available at https://github.com/SchwarzEM/ems_perl/tree/master/gff.

### Independently predicting genes in our VC2010 assembly

In the VC2010 assembly, we predicted protein-coding gene structures with AUGUSTUS 3.3 (Stanke et al. 2008), using *C. elegans*-specific gene parameters and with hints generated by BLAT-aligning protein-coding DNA sequences (CDS DNA) from the N2 reference assembly to the VC2010 assembly; the arguments used were ‘*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=$HOME/src/augustus-3.3/config/extrinsic/extrinsic.ME.cfg progress=true gff3=on*’. We predicted noncoding-RNA genes with cmsearch from Infernal 1.1.2 (Nawrocki and Eddy 2013) and Rfam 13.0 (Kalvari et al. 2018); using the argument ‘*--cut_ga*’ to invoke model-specific reporting thresholds. For subsequent overlap analysis, we produced a BED6 annotation file from the original Infernal/Rfam output table with rfam_to_bed6.pl (available at https://github.com/SchwarzEM/ems_perl/tree/master/gff) and reformatted it to GFF3 with the bed_to_gff3 program of GenomeTools (Gremme et al. 2013), downloaded from http://genometools.org/pub/binary_distributions/gt-1.5.10-Linux_i386-64bit.tar.gz.

### Analyzing VC2010-assembly-specific genes and tandem repeat regions

To characterize possible new VC2010-assembly-specific genes, we first checked them for overlapping spans with lifted-over N2 genes by running the intersect program in BEDTools with the arguments ‘*-loj -s*’ and ‘*-loj -S*’. Note that these arguments check for span overlaps with same-strand and opposite-strand genes, respectively; such overlaps do not automatically mean that the genes are actually equivalent (for instance, a smaller gene residing entirely within the intron of a large gene would nevertheless be scored as overlapping the larger gene). The main purpose of such analysis, when coupled with BLASTP and BLASTN scores, was to detect possible mispredictions of alternative/VC2010-assembly-specific exons as free-standing genes.

Having tested overlaps with N2, we characterized possible new VC2010-assembly-specific genes for identity or similarity to known genes in *C. elegans* N2 at both the protein and the DNA level, by using BLASTP and BLASTN (respectively) against predicted protein products (downloaded from *ftp://ftp.wormbase.org/pub/wormbase/releases/WS264/species/c_elegans/PRJNA13758/c_elegans.PRJNA13758.WS264.protein.fa.gz*) or CDS DNA sequences (downloaded from *ftp://ftp.wormbase.org/pub/wormbase/releases/WS264/species/c_elegans/PRJNA13758/c_elegans.PRJNA13758.WS264.CDS_transcripts.fa.gz*) of the N2 genes. Both BLASTP and BLASTN were from BLAST 2.7.1. For BLASTP, we used the arguments ‘*-outfmt 7 -max_target_seqs 1 -max_hsps 1 -evalue 0.1*’; for BLASTN, we used the similar arguments ‘*dust no -outfmt 7 -max_target_seqs 1 -max_hsps 1 -evalue 0.1*’. As a control for failure to detect proteins in N2 that might be present in a more fully assembled genome, we also carried out BLASTP against the proteome of *Caenorhabditis nigoni* (downloaded from *ftp://ftp.wormbase.org/pub/wormbase/releases/WS264/species/c_nigoni/PRJNA384657/c_nigoni.PRJNA384657.WS264.protein.fa.gz*), a recently published PacBio-based genome of a species closely related to *C. elegans* (Yin et al. 2018).

To characterize basic properties of the possible new gene products independently of BLAST, we predicted signal sequences and transmembrane sequences with Phobius 1.01 (Käll et al. 2004), coiled-coils with NCoils (Lupas 1996), low-complexity domains with PSEG (Wootton 1994), and protein motifs from the Pfam31 database with hmmscan in HMMER 3.1b2 (Eddy 2009; Finn et al. 2016), using the argument ‘*--cut_ga*’ to impose family-specific significance thresholds. To simplify the above analyses, we used only the longest isoform of each gene (both for VC2010 and for N2 or *C. nigoni*). Longest isoforms were extracted from proteomes with get_largest_isoforms.pl (available at https://github.com/SchwarzEM/ems_perl/tree/master/fasta).

To determine which AUGUSTUS genes shared identical coding residues with lifted-over N2 genes, we used the Perl script diff_gff3_genesets.pl (available at https://github.com/SchwarzEM/ems_perl/tree/master/gff). For all other tests of overlap, we used the intersect program of BEDTools 2.27.1 to identify which genes fell completely within VC2010-assembly-specific genome regions. Further details are given in Supplemental Methods.

Comparisons of tandem repeat expansions in VC2010 to YAC-derived genomic regions of the N2 genome assembly, as with comparisons of N2-encoded to VC2010-encoded genes, relied on annotation liftovers followed by genome-interval intersections via BEDTools. Further details are given in Supplemental Methods.

Regions of the VC2010 assembly and their genome annotations (Fig. 4) were visualized with JBrowse 1.16.2 (Skinner et al. 2009).

## Data access

All raw sequencing reads from this study have been submitted to NCBI BioProject database (BioProject; https://www.ncbi.nlm.nih.gov/bioproject/) under accession numbers PRJNA430756, PRJNA482888, and PRJNA482889. The genome assembly from this study has been submitted to the European Nucleotide Archive (ENA; https://www.ebi.ac.uk/ena) under accession number PRJEB28388. Precursor VC2010 genome assemblies have been archived in OSF (https://osf.io/bscjx; doi:10.17605/osf.io/bscjx).

## Acknowledgments

We thank Koichiro Doi, Tân Vũ, and Kosuke Kudo for developing a tool for visualizing dot plots with alignments of reads and genes; Michael McGurk for discussions on noncoding repetitive genomic DNA; Robert Waterston and Erik Andersen for discussions on genetic diversity within N2; Michael Paulini from WormBase for managing genome assembly submission to ENA; Roberto Gomez for suggesting using the NCBI fork of ACeDB; and Titus Brown and the Michigan State University High-Performance Computing Center (supported by U.S. Department of Agriculture grant 2010-65205-20361 and NIFA–National Science Foundation (NSF) grant IOS-0923812) for computational support. Additional computing was enabled by a start-up allocation from NSF XSEDE (TG-MCB180039). This study was supported by the Japan Science and Technology Corporation (CREST, JPMJCR13W3) and AMED (Japan Agency for Medical Research and Development, GRIFIN) grants to S.M.; National Institutes of Health (NIH) grant GM37706/GM130366 to A.Z.F.; NIH grant AI111173, Moore Foundation Grant No. 4551, and Cornell University start-up funds to E.M.S.; NIH grant OD010440 to A.E.R.; an Arnold O. Beckman Postdoctoral Fellowship and a Stanford Medicine Dean's Fellowship to M.J.S.; and a Helen Hay Whitney Postdoctoral Fellowship to L.W.

*Author contributions:* J.Y. assembled and polished genomes, closed gaps, assessed genome quality, and analyzed genome structures; K.I. analyzed repetitive elements and tandem repeats; M.J.S. developed protocols for long/ultra-long genomic DNA purification and sequencing, carried them out, and did Illumina MiSeq sequencing; K.L.A. generated and archived clonal *C. elegans* strains and carried out long/ultra-long genomic DNA purification and sequencing; I.G. generated clonal *C. elegans* strains, identified genomic structural variants, and did Illumina sequencing; L.W. did Illumina MiSeq sequencing; C.L.S. developed protocols for long/ultra-long genomic DNA purification and sequencing; M.L.E. generated clonal *C. elegans* strains; A.E.R. helped conceive this study, archived clonal *C. elegans* strains, and coordinated work; A.Z.F. helped conceive this study, led discussions of specific and general approaches toward development of a high-quality genome, and coordinated work; S.M. coded software, analyzed data, reviewed other code and data analyses, and coordinated work; E.M.S. devised strategies for genome assembly, lifted over pre-existing genome annotations, predicted novel genes, and coordinated work. All authors contributed to the manuscript, with most writing by A.Z.F., S.M., and E.M.S.

## 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

Azzalin CM, Reichenbach P, Khoriauli L, Giulotto E, Lingner J. 2007. Telomeric repeat containing RNA and RNA surveillance factors at mammalian chromosome ends. *Science* **318:** 798–801. doi:10.1126/science.1147182

Benson G. 1999. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res* **27:** 573–580. doi:10.1093/nar/27.2.573

Berlin K, Koren S, Chin C-S, Drake JP, Landolin JM, Phillippy AM. 2015. Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. *Nat Biotechnol* **33:** 623–630. doi:10.1038/nbt.3238

Bessereau JL. 2006. Transposons in *C. elegans*. *WormBook* 1–13. doi:10.1895/wormbook.1.70.1

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

Burke M, Scholl EH, Bird DM, Schaff JE, Colman SD, Crowell R, Diener S, Gordon O, Graham S, Wang X, et al. 2015. The plant parasite *Pratylenchus coffeae* carries a minimal nematode genome. *Nematology* **17:** 621–637. doi:10.1163/15685411-00002901

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

Chaisson MJ, Tesler G. 2012. Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. *BMC Bioinformatics* **13:** 238. doi:10.1186/1471-2105-13-238

Chaisson MJP, Huddleston J, Dennis MY, Sudmant PH, Malig M, Hormozdiari F, Antonacci F, Surti U, Sandstrom R, Boitano M, et al. 2015. Resolving the complexity of the human genome using single-molecule sequencing. *Nature* **517:** 608–611. doi:10.1038/nature13907

Chakraborty M, VanKuren NW, Zhao R, Zhang X, Kalsow S, Emerson JJ. 2018. Hidden genetic variation shapes the structure of functional elements in *Drosophila*. *Nat Genet* **50:** 20–25. doi:10.1038/s41588-017-0010-y

Chang CH, Larracuente AM. 2019. Heterochromatin-enriched assemblies reveal the sequence and organization of the *Drosophila melanogaster* Y chromosome. *Genetics* **211:** 333–348. doi:10.1534/genetics.118.301765

Charlesworth B, Langley CH, Stephan W. 1986. The evolution of restricted recombination and the accumulation of repeated DNA sequences. *Genetics* **112:** 947–962.

Chin CS, Alexander DH, Marks P, Klammer A, Drake J, Heiner C, Clum A, Copeland A, Huddleston J, Eichler EE, et al. 2013. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat Methods* **10:** 563–569. doi:10.1038/nmeth.2474

Chin CS, Peluso P, Sedlazeck FJ, Nattestad M, Concepcion GT, Clum A, Dunn C, O'Malley R, Figueroa-Balderas R, Morales-Cruz A, et al. 2016. Phased diploid genome assembly with single-molecule real-time sequencing. *Nat Methods* **13:** 1050–1054. doi:10.1038/nmeth.4035

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

Coulson A, Waterston R, Kiff J, Sulston J, Kohara Y. 1988. Genome linking with yeast artificial chromosomes. *Nature* **335:** 184–186. doi:10.1038/335184a0

Coulson A, Kozono Y, Lutterbach B, Shownkeen R, Sulston J, Waterston R. 1991. YACs and the *C. elegans* genome. *Bioessays* **13:** 413–417. doi:10.1002/bies.950130809

Coulson A, Huynh C, Kozono Y, Shownkeen R. 1995. The physical map of the *Caenorhabditis elegans* genome. *Methods Cell Biol* **48:** 533–550. doi:10.1016/S0091-679X(08)61402-8

Deamer D, Akeson M, Branton D. 2016. Three decades of nanopore sequencing. *Nat Biotechnol* **34:** 518–524. doi:10.1038/nbt.3423

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

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

Eccles D, Chandler J, Camberis M, Henrissat B, Koren S, Le Gros G, Ewbank JJ. 2018. De novo assembly of the complex genome of *Nippostrongylus brasiliensis* using MinION long reads. *BMC Biol* **16:** 6. doi:10.1186/s12915-017-0473-4

Eddy SR. 2009. A new generation of homology search tools based on probabilistic inference. *Genome Inform* **23:** 205–211.

Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, Peluso P, Rank D, Baybayan P, Bettman B, et al. 2009. Real-time DNA sequencing from single polymerase molecules. *Science* **323:** 133–138. doi:10.1126/science.1162986

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

Finn RD, Coggill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, Potter SC, Punta M, Qureshi M, Sangrador-Vegas A, et al. 2016. The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res* **44:** D279–D285. doi:10.1093/nar/gkv1344

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

Frezal L, Felix MA. 2015. *C. elegans* outside the Petri dish. *eLife* **4:** e05849. doi:10.7554/eLife.05849

Friedman S, Freitag M. 2017. Evolving centromeres and kinetochores. *Adv Genet* **98:** 1–41. doi:10.1016/bs.adgen.2017.07.001

Fu L, Niu B, Zhu Z, Wu S, Li W. 2012. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28:** 3150–3152. doi:10.1093/bioinformatics/bts565

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

Godiska R, Mead D, Dhodda V, Wu C, Hochstein R, Karsi A, Usdin K, Entezam A, Ravin N. 2010. Linear plasmid vector for cloning of repetitive or unstable sequences in *Escherichia coli*. *Nucleic Acids Res* **38:** e88. doi:10.1093/nar/gkp1181

Goodwin S, Gurtowski J, Ethe-Sayers S, Deshpande P, Schatz MC, McCombie WR. 2015. Oxford Nanopore sequencing, hybrid error correction, and de novo assembly of a eukaryotic genome. *Genome Res* **25:** 1750–1756. doi:10.1101/gr.191395.115

Gordon D, Huddleston J, Chaisson MJP, Hill CM, Kronenberg ZN, Munson KM, Malig M, Raja A, Fiddes I, Hillier LW, et al. 2016. Long-read sequence assembly of the gorilla genome. *Science* **352:** aae0344. doi:10.1126/science.aae0344

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

Gurevich A, Saveliev V, Vyahhi N, Tesler G. 2013. QUAST: quality assessment tool for genome assemblies. *Bioinformatics* **29:** 1072–1075. doi:10.1093/bioinformatics/btt086

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

Ichikawa K, Tomioka S, Suzuki Y, Nakamura R, Doi K, Yoshimura J, Kumagai M, Inoue Y, Uchida Y, Irie N, et al. 2017. Centromere evolution and CpG methylation during vertebrate speciation. *Nat Commun* **8:** 1833. doi:10.1038/s41467-017-01982-7

Jain M, Koren S, Miga KH, Quick J, Rand AC, Sasani TA, Tyson JR, Beggs AD, Dilthey AT, Fiddes IT, et al. 2018a. Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nat Biotechnol* **36:** 338–345. doi:10.1038/nbt.4060

Jain M, Olsen HE, Turner DJ, Stoddart D, Bulazel KV, Paten B, Haussler D, Willard HF, Akeson M, Miga KH. 2018b. Linear assembly of a human centromere on the Y chromosome. *Nat Biotechnol* **36:** 321–323. doi:10.1038/nbt.4109

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

Kalvari I, Argasinska J, Quinones-Olvera N, Nawrocki EP, Rivas E, Eddy SR, Bateman A, Finn RD, Petrov AI. 2018. Rfam 13.0: shifting to a genome-centric resource for non-coding RNA families. *Nucleic Acids Res* **46:** D335–D342. doi:10.1093/nar/gkx1038

Kamath GM, Shomorony I, Xia F, Courtade TA, Tse DN. 2017. HINGE: long-read assembly achieves optimal repeat resolution. *Genome Res* **27:** 747–756. doi:10.1101/gr.216465.116

Kasahara M, Morishita S. 2006. *Large-scale genome sequence processing*. World Scientific, Singapore.

Kent WJ, Baertsch R, Hinrichs A, Miller W, Haussler D. 2003. Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes. *Proc Natl Acad Sci* **100:** 11484–11489. doi:10.1073/pnas.1932072100

Kim SH, Hwang SB, Chung IK, Lee J. 2003. Sequence-specific binding to telomeric DNA by CEH-37, a homeodomain protein in the nematode *Caenorhabditis elegans*. *J Biol Chem* **278:** 28038–28044. doi:10.1074/jbc.M302192200

Koren S, Schatz MC, Walenz BP, Martin J, Howard JT, Ganapathy G, Wang Z, Rasko DA, McCombie WR, Jarvis ED, et al. 2012. Hybrid error correction and *de novo* assembly of single-molecule sequencing reads. *Nat Biotechnol* **30:** 693–700. doi:10.1038/nbt.2280

Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. 2017. Canu: scalable and accurate long-read assembly via adaptive *k*-mer weighting and repeat separation. *Genome Res* **27:** 722–736. doi:10.1101/gr.215087.116

Korhonen PK, Young ND, Gasser RB. 2016. Making sense of genomes of parasitic worms: tackling bioinformatic challenges. *Biotechnol Adv* **34:** 663–686. doi:10.1016/j.biotechadv.2016.03.001

Korlach J, Marks PJ, Cicero RL, Gray JJ, Murphy DL, Roitman DB, Pham TT, Otto G, Foquet M, Turner SW. 2008. Selective aluminum passivation for targeted immobilization of single DNA polymerase molecules in zero-mode waveguide nanostructures. *Proc Natl Acad Sci* **105:** 1176–1181. doi:10.1073/pnas.0710982105

Kronenberg ZN, Fiddes IT, Gordon D, Murali S, Cantsilieris S, Meyerson OS, Underwood JG, Nelson BJ, Chaisson MJP, Dougherty ML, et al. 2018. High-resolution comparative analysis of great ape genomes. *Science* **360:** eaar6343. doi:10.1126/science.aar6343

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 RYN, Howe KL, Harris TW, Arnaboldi V, Cain S, Chan J, Chen WJ, Davis P, Gao S, Grove C, et al. 2018. WormBase 2017: molting into a new stage. *Nucleic Acids Res* **46:** D869–D874. doi:10.1093/nar/gkx998

Levene MJ, Korlach J, Turner SW, Foquet M, Craighead HG, Webb WW. 2003. Zero-mode waveguides for single-molecule analysis at high concentrations. *Science* **299:** 682–686. doi:10.1126/science.1079700

Li H. 2016. Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences. *Bioinformatics* **32:** 2103–2110. doi:10.1093/bioinformatics/btw152

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

Loman NJ, Quinlan AR. 2014. Poretools: a toolkit for analyzing nanopore sequence data. *Bioinformatics* **30:** 3399–3401. doi:10.1093/bioinformatics/btu555

Loman NJ, Watson M. 2015. Successful test launch for nanopore sequencing. *Nat Methods* **12:** 303–304. doi:10.1038/nmeth.3327

Loman NJ, Quick J, Simpson JT. 2015. A complete bacterial genome assembled *de novo* using only nanopore sequencing data. *Nat Methods* **12:** 733–735. doi:10.1038/nmeth.3444

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

Lynch M. 2007. The frailty of adaptive hypotheses for the origins of organismal complexity. *Proc Natl Acad Sci* **104:** 8597–8604. doi:10.1073/pnas.0702207104

McCaffrey J, Young E, Lassahn K, Sibert J, Pastor S, Riethman H, Xiao M. 2017. High-throughput single-molecule telomere characterization. *Genome Res* **27:** 1904–1915. doi:10.1101/gr.222422.117

Myers G. 2014. Efficient local alignment discovery amongst noisy long reads. In *Algorithms in bioinformatics. WABI 2014. Lecture Notes in Computer Science*, Vol. 8701 (ed. Brown D, Morgenstern B), pp. 52–67. Springer, Berlin.

Nattestad M, Chin CS, Schatz M. 2016. Ribbon: visualizing complex genome alignments and structural variation. bioRxiv doi:10.1101/082123

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

Nicorici D, Satalan M, Edgren H, Kangaspeska S, Murumagi A, Kallioniemi O, Virtanen S, Kilkku O. 2014. FusionCatcher – a tool for finding somatic fusion genes in paired-end RNA-sequencing data. bioRxiv doi:10.1101/011650

Nowoshilow S, Schloissnig S, Fei JF, Dahl A, Pang AWC, Pippel M, Winkler S, Hastie AR, Young G, Roscito JG, et al. 2018. The axolotl genome and the evolution of key tissue formation regulators. *Nature* **554:** 50–55. doi:10.1038/nature25458

Pendleton M, Sebra R, Pang AWC, Ummat A, Franzen O, Rausch T, Stütz AM, Stedman W, Anantharaman T, Hastie A, et al. 2015. Assembly and diploid architecture of an individual human genome via single-molecule technologies. *Nat Methods* **12:** 780–786. doi:10.1038/nmeth.3454

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

Raffaele S, Kamoun S. 2012. Genome evolution in filamentous plant pathogens: why bigger can be better. *Nat Rev Microbiol* **10:** 417–430. doi:10.1038/nrmicro2790

Richardson SM, Mitchell LA, Stracquadanio G, Yang K, Dymond JS, DiCarlo JE, Lee D, Huang CL, Chandrasegaran S, Cai Y, et al. 2017. Design of a synthetic yeast genome. *Science* **355:** 1040–1044. doi:10.1126/science.aaf4557

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

Ross MG, Russ C, Costello M, Hollinger A, Lennon NJ, Hegarty R, Nusbaum C, Jaffe DB. 2013. Characterizing and measuring bias in sequence data. *Genome Biol* **14:** R51. doi:10.1186/gb-2013-14-5-r51

Schoeftner S, Blasco MA. 2008. Developmentally regulated transcription of mammalian telomeres by DNA-dependent RNA polymerase II. *Nat Cell Biol* **10:** 228–236. doi:10.1038/ncb1685

Schwartz DC, Cantor CR. 1984. Separation of yeast chromosome-sized DNAs by pulsed field gradient gel electrophoresis. *Cell* **37:** 67–75. doi:10.1016/0092-8674(84)90301-5

Sha K, Gu SG, Pantalena-Filho LC, Goh A, Fleenor J, Blanchard D, Krishna C, Fire A. 2010. Distributed probing of chromatin structure *in vivo* reveals pervasive chromatin accessibility for expressed and non-expressed genes during tissue differentiation in *C. elegans*. *BMC Genomics* **11:** 465. doi:10.1186/1471-2164-11-465

Shoura MJ, Gabdank I, Hansen L, Merker J, Gotlib J, Levene SD, Fire AZ. 2017. Intricate and cell type-specific populations of endogenous circular DNA (eccDNA) in *Caenorhabditis elegans* and *Homo sapiens*. *G3 (Bethesda)* **7:** 3295–3303. doi:10.1534/g3.117.300141

Skinner ME, Uzilov AV, Stein LD, Mungall CJ, Holmes IH. 2009. JBrowse: a next-generation genome browser. *Genome Res* **19:** 1630–1638. doi:10.1101/gr.094607.109

Solares EA, Chakraborty M, Miller DE, Kalsow S, Hall K, Perera AG, Emerson JJ, Hawley RS. 2018. Rapid low-cost assembly of the *Drosophila melanogaster* reference genome using low-coverage, long-read sequencing. *G3 (Bethesda)* **8:** 3143–3154. doi:10.1534/g3.118.200162

Speir ML, Zweig AS, Rosenbloom KR, Raney BJ, Paten B, Nejad P, Lee BT, Learned K, Karolchik D, Hinrichs AS, et al. 2016. The UCSC Genome Browser database: 2016 update. *Nucleic Acids Res* **44:** D717–D725. doi:10.1093/nar/gkv1275

Spieth J, Lawson D, Davis P, Williams G, Howe K. 2014. Overview of gene structure in *C. elegans*. *WormBook* 1–18. doi:10.1895/wormbook.1.65.1

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

Stricklin SL, Griffiths-Jones S, Eddy SR. 2005. *C. elegans* noncoding RNA genes. *WormBook* 1–7. doi:10.1895/wormbook.1.1.1

Subirana JA, Messeguer X. 2013. A satellite explosion in the genome of holocentric nematodes. *PLoS One* **8:** e62221. doi:10.1371/journal.pone.0062221

Sulston JE, Brenner S. 1974. The DNA of *Caenorhabditis elegans*. *Genetics* **77:** 95–104.

Swan KA, Curtis DE, McKusick KB, Voinov AV, Mapa FA, Cancilla MR. 2002. High-throughput gene mapping in *Caenorhabditis elegans*. *Genome Res* **12:** 1100–1105. doi:10.1101/gr.208902

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

VanBuren R, Bryant D, Edger PP, Tang H, Burgess D, Challabathula D, Spittle K, Hall R, Gu J, Lyons E, et al. 2015. Single-molecule sequencing of the desiccation-tolerant grass *Oropetium thomaeum*. *Nature* **527:** 508–511. doi:10.1038/nature15714

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

Vollger MR, Dishuck PC, Sorensen M, Welch AE, Dang V, Dougherty ML, Graves-Lindsay TA, Wilson RK, Chaisson MJP, Eichler EE. 2019. Long-read sequence and assembly of segmental duplications. *Nat Methods* **16:** 88–94. doi:10.1038/s41592-018-0236-3

Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, Cuomo CA, Zeng Q, Wortman J, Young SK, et al. 2014. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* **9:** e112963. doi:10.1371/journal.pone.0112963

Waterhouse RM, Seppey M, Simão FA, Manni M, Ioannidis P, Klioutchnikov G, Kriventseva EV, Zdobnov EM. 2018. BUSCO applications from quality assessments to gene prediction and phylogenomics. *Mol Biol Evol* **35:** 543–548. doi:10.1093/molbev/msx319

Wicks SR, Yeh RT, Gish WR, Waterston RH, Plasterk RH. 2001. Rapid gene mapping in *Caenorhabditis elegans* using a high density polymorphism map. *Nat Genet* **28:** 160–164. doi:10.1038/88878

Wicky C, Villeneuve AM, Lauper N, Codourey L, Tobler H, Muller F. 1996. Telomeric repeats (TTAGGC)$_n$ are sufficient for chromosome capping function in *Caenorhabditis elegans*. *Proc Natl Acad Sci* **93:** 8983–8988. doi:10.1073/pnas.93.17.8983

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

Yin D, Schwarz EM, Thomas CG, Felde RL, Korf IF, Cutter AD, Schartner CM, Ralston EJ, Meyer BJ, Haag ES. 2018. Rapid genome shrinkage in a self-fertile nematode reveals sperm competition proteins. *Science* **359:** 55–61. doi:10.1126/science.aao0827

*Received October 4, 2018; accepted in revised form March 11, 2019.*