Recompleting the *Caenorhabditis elegans* genome [Supplemental Text]

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, Stanford University, Stanford, CA 94305, USA. [3]Department of Genetics, Stanford University, Stanford, CA 94305, USA. [4]Department of Zoology and Michael Smith Laboratories, University of British Columbia, Vancouver, 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, NY 14853, USA. [7]These authors contributed equally. [*]Correspondence: *afire@stanford.edu*; *moris@edu.k.u-tokyo.ac.jp*; *ems394@cornell.edu*.

SUPPLEMENTAL FIGURE LEGENDS

Supplemental Figure S1: Reassembling long sequencing reads

By aligning long reads with regions of the VC2010 assembly that had large-scale differences from the N2 assembly, and then locally reassembling those long reads, we could correct 20 erroneous regions in the VC2010 assembly. In each panel, the first row describes the revised VC2010 contig with its assembler, read set and identifier, and corrected region. Below the description, we show alignments of raw reads that suggest errors of the VC2010 contig. We revised the erroneous VC2010 contig by assembling all raw reads anchored to the erroneous region into a corrected contig with Canu; this reassembly did not use any information from the N2 assembly. Below the alignments, we display a dot plot for part of the provisional VC2010 assembly around the erroneous region before revision (in the y-axis) and its corresponding region in the N2 reference assembly (in the x-axis). Another dot plot below is one between the revised contig (in the y-axis) and its corresponding N2 region (in the x-axis), confirming that 14 locally reassembled contigs matched the N2 reference, while we observed that the remaining six contigs had different patterns of tandem repeats.

Supplemental Figure S2: Contigs aligned to the N2 reference assembly

Six panels display the layout of contigs in six chromosomes of the N2 assembly. The respective red and blue thick lines show alignments of contigs in the plus and minus strands. The label beside each contig line shows the identifier of the contig.

Supplemental Figure S3: Filling gaps between contigs generated with PacBio reads

Each panel shows how we filled gaps between contigs that we generated from PacBio reads using Canu. Contigs are highlighted orange, with their positions in the VC2010 assembly shown in the top table of each figure. Other sequences for gap filling are colored differently depending on their types. Contigs assembled by FALCON, miniasm, or HINGE are yellow, contigs assembled from reads around each gap are green (see Supplemental Figure S1), tandem repeat expansions are blue, and corrected reads are light blue. Contigs subsumed by longer contigs are colored white, and misassembled contigs are black. The remaining gaps are black and their surrounding tandem repeats are gray. Below the top table, we show a schematic figure to illustrate the steps of filling gaps. We connected two contigs if they shared an identical subsequence of ≧100 kb at their ends. After connecting two contigs approximately, we reconfirmed that two neighboring sequences were truly overlapping at nucleotide resolution by making sure that multiple alignments of reads at the boundary of two sequences had overlaps of at least 100 nt.

Supplemental Figure S4: Closing or partially closing five gaps in the VC2010 assembly

(A) Closing the gap in Chromosome I (nt 4,605,602-4,668,785). We measured the similarity among repeat unit occurrences in an observed tandem repeat as the match ratio of an optimal local alignment between the observed tandem repeat and an ideal tandem repeat with no mismatches (a series of repeat unit copies). We determined the unit string using PacBio contig #0. Afterwards, we decided the number of repeat unit occurrences using the read with the largest number of occurrences, which is colored red in the table. **(B)** Closing the gap in Chromosome II (nt 14,525,986-14,566,400). Two Nanopore reads #1 and #2 around the gap were different in terms of number of both repeat unit occurrences and similarity among units. We determined the respective numbers of units in tandem repeats A and B from reads #1 and #2 that had more units in them. **(C)** Closing the gap in Chromosome X (nt 4,132,271-4,153,270). Although no reads covered both tandem repeats A and B simultaneously, respective tandem repeats were spanned by reads #1 and #2 that overlapped and shared the two tandem repeats in common. We therefore filled the gap using the information of unit numbers in the two reads. **(D)** Partially closing the gap in Chromosome X (nt 5,170,258-5,170,357). Tandem repeat A was spanned by read #1, and its number of unit occurrences was estimated to be 98 from read #1. By contrast, tandem repeat B was covered partially by reads #1 and #2, and a lower bound of unit number, 1697, was estimated from read #2. **(E)** Partially closing the gap in Chromosome III (nt 7,587,315-7,682,900). Tandem repeats A and B in two PacBio contigs #0 are truncated and underestimated.

However, no Nanopore reads span each tandem repeat A and B, and we estimated a lower bound for the number of units from reads #2 and #3. **(F)** Discrepancy between reads called by two Nanopore basecallers, MinKNOW and Albacore v. 2.1.7. The dot plots show two reads of an identical DNA fragment originating from Chromosome I. We observed large differences between them in terms of read length, number of units, and similarity among units.

Supplemental Figure S5: Comparisons of N2 to VC2010 for five gaps

For each of five regions with large gaps in the reference N2 assembly and its corresponding region in the VC2010 assembly, we show a dot plot between the N2 reference assembly in the x-axis and its matching VC2010 assembly region in the y-axis.

Supplemental Figure S6: Dot plots for 109 large tandem repeat regions

Dot plots are shown for 109 large tandem repeat expansions of >1 kb that have nearly identical units but differ in the number of unit occurrences between the N2 and VC2010 assemblies. In each dot plot, the respective x-axis and y-axis show the N2 reference assembly versus our VC2010 assembly, with the range in N2 displayed at the top. Blue lines right to each dot plot represent alignments of VC2010 reads to our VC2010 assembly in the y-axis. We checked whether or not each tandem repeat in our VC2010 assembly was covered by raw PacBio reads; results are listed in column N of Supplemental Table S9 (as "Yes", "Partially", or "No").

Supplemental Figure S7: Comparing the N2 and VC2010 references for imperfect tandem repeats

Each of 108 figures shows a dot plot between a imperfect tandem repeat in the VC2010 reference (y-axis) and its best matching region in the N2 reference (x-axis).

Supplemental Figure S8: Dot plots for six pairs of telomeric regions (N2 versus VC2010)

Dot plots are shown for six pairs of telomeric regions of our VC2010 assembly (in the x-axis) and the N2 reference (in the y-axis) at the left (5') and right (3') ends of Chromosomes I, II, III, IV, V, and X. We observed that telomeric tandem repeats of unit GCCTAA (or its reverse complement TTAGGC) in our VC2010 assembly were longer than those in the N2 assembly, and were at the ends of all chromosomes except for the right ends of Chromosome I and III. We could not sequence these two right ends through their telomeric tandem repeats. We therefore attempted to extend the right ends using

raw Nanopore and PacBio reads (see Supplemental Figure S9). Supplemental Table S14 shows a precise comparison between telomeric tandem repeats between the VC2010 and N2 assemblies.

Supplemental Figure S9: Matching raw long reads to the right ends of Chromosomes I and III

We searched for raw PacBio and Nanopore reads that matched the right (3') ends of Chromosomes I and III in the VC2010 assembly. **(A)** One Nanopore read (Nanopore 1) matched the right end of Chromosome I. The read had eight tandem copies of a unit and one partial copy of the unit (blue). Due to a high error rate of Nanopore reads, it is difficult to determine the string of the repeat unit; however, its size is estimated to be about 7 kb. We also found another Nanopore read (Nanopore 2) that had five tandem unit copies (blue) and a telomeric tandem repeat (red) at the tail (see a precise analysis in Supplemental Table S14). The schematic figure at the top illustrates a candidate multiple alignment of the 3'-end VC2010 region and the two Nanopore reads. The lower six dot plots show matches between all pairs of the three sequences. Although there is ambiguity in aligning the two Nanopore reads, we certainly observe the presence of a subtelomeric tandem repeat in the proximity of a telomeric GCCTAA tandem repeat. **(B)** One Nanopore read matched the right end of Chromosome III. Although the Nanopore read appeared to have many errors, it had subtelomeric tandem repeat units (blue) and a telomeric repeat (red) at the tail, confirming a telomere structure at the right end of Chromosome III. We also found another PacBio read with a shorter subtelomeric repeat (blue) but a longer telomeric tandem repeat. We show a schematic multiple alignment of the three sequences at the top, and confirm the alignment using the six dot plots of the three sequences below.

Supplemental Figure S10: Dot plots for 100 apparent insertions of >100 nt into VC2010

We identified 100 candidate insertions of >100 nt into the VC2010 assembly with respect to the N2 assembly; we then compared these insertions to the genome assemblies of two additional strains, PD2182 and PD2183, that are outgroups of N2 and VC2010. Upon comparison, most of the apparent insertions proved to be deficiencies in the N2 reference assembly. The left figure in each page displays a dot plot of two reciprocally best matching regions between the VC2010 assembly in the x-axis and the N2 reference in the y-axis, while the respective middle and right figures show dot plots of regions between the VC2010 assembly (x-axis) and PD2182 (y-axis), and between the VC2010 assembly and PD2183. In each dot plot, a range of the VC2010 assembly is shown at the top, and a candidate insertion into the VC2010 is displayed at 10 kb from the left end of the range. In some cases, we identified altered gene structures by aligning the WormBase gene set (release WS264; Lee et al. 2018) to the VC2010

assembly. When genes are aligned with the VC2010 assembly, we show their exon-intron structures below the x-axis. Thin blue lines below the x-axis depict alignments of long reads that map to the VC2010 region and confirm the presence of the insertion in the VC2010 region. Similarly, thin blue lines next to the y-axis in the left (middle, right, respectively) dot plot indicate alignments of VC2010 (or PD2182 or PD2183) reads to the N2 (or PD2182 or PD2183) genomic region around the insertions. In the first page, for example, the left dot plot has no alignments of VC2010 reads to the N2 reference next to the y-axis because the candidate 121-nt insertion is missing in the N2 reference. By contrast, the diagonal line in the middle (right, respectively) dot plot show the agreement of VC2010 assembly and PD2182 (PD2183) genomes in the region, implying the insertion is an error in the N2 reference, which is also supported by blue lines next to the y-axis. Of 100 candidate insertions into VC2010 with respect N2, 89 were confirmed as present in VC2010, while the remaining 10 were undetermined because the corresponding regions in PD2182 and PD2183 matched neither VC2010 nor N2 (Supplemental Table S16).

Supplemental Figure S11: Characterizing six apparent deletions in VC2010

We compared the N2, VC2010, PD2182, and PD2183 genome assemblies, and identified six apparent deletions in VC2010. We then assigned each deletion to one of three classes: a real deletion that occurred in the lineage from N2 to VC2010; an assembly error in the N2 genome; or an undetermined variant. As in Supplemental Figure S10, we show dot plots with alignments of genes and raw PacBio reads, and in some cases, identified altered gene structures by aligning WormBase WS264 genes to VC2010. The first page, for example, shows a deletion in the lineage from N2 to VC2010 because the deletion is not observed in N2, PD2182, or PD2183, and also because the PD2182 and PD2183 genome assemblies in this region are validated by the alignments of raw PacBio reads. The third page presents a case where the deletion is an error in the N2 genome assembly because the VC2010, PD2182, and PD2183 genomic regions are identical. Of six possible deletions, four occurred in the lineage from N2 to VC2010, one was an assembly error or mutation in the N2 reference, and one was undetermined because of inconsistent assemblies (Supplemental Table S17).

Supplemental Figure S12: Characterizing 30 apparent genomic duplications in N2 or VC2010

We identified 30 possible genomic duplications in N2 or VC2010 by four-way comparison of the N2, VC2010, PD2182, and PD2183 genome assemblies, and assigned each duplication to one of three classes by analyzing dot plots as in Supplemental Figures S10-S11; also as before, we identified altered

gene structures by aligning WormBase WS264 genes to our VC2010 assembly. The first page presents a genomic duplication that we infer to have taken place in the lineage from N2 to VC2010, because the N2, PD2182, and PD2183 assemblies agree. Of 30 possible duplications, 14 occurred in the lineage from N2 to VC2010 and are thus VC2010-specific; seven were assembly errors or mutations in the N2 reference; and 9 were undetermined because of inconsistent assemblies (Supplemental Table S18).

SUPPLEMENTAL TABLE LEGENDS

Supplemental Table S1: Raw genome sequencing reads

We used PacBio RS II and Oxford Nanopore Technology MinION machines to sequence long reads from three clonal VC2010-derived strains PD1073, PD1074, and PD1075; we used these long reads, in turn, to assemble the VC2010 genome. The table shows the statistics of these raw sequencing reads. We did not use reads from our first Nanopore run because this run generated much shorter reads than subsequent runs did. The last column (column P) shows the N*xx* Nanopore read length, where *xx* ranges from 10% to 99%, among all Nanopore reads collected in the second to seventh runs.

Supplemental Table S2: Seven initial genome assemblies

We used two read sets, one set with PacBio reads only and the other with PacBio and Nanopore reads. We then used four genome assemblies for long read sequencing (Canu, miniasm, FALCON, and HINGE) to assemble the two read sets into seven contig sets. We used HINGE for assembling the PacBio and Nanopore read set to close gaps. After generating contigs, we polished bases in the contigs of all assemblies using Quiver. The table shows the statistics of contigs in the seven assemblies. The seven assembly sequences themselves are provided in a permanent data archive at the Open Science Framework (*https://osf.io/jx89y*; doi:10.17605/osf.io/jx89y).

Supplemental Table S3: Covered assembly gaps

We used the Canu genome assembly as the primary one and attempted to fill its gaps with contigs in the other genome assemblies. The first column shows chromosome numbers. The second column displays the primary contigs in the Canu assembly, which are colored orange, and rows between orange contigs represent gaps. The third column shows that most of gaps are closed by contigs (colored yellow) in the other six genome assemblies, corrected reads (light blue), or extremely long Nanopore reads (red). The two remaining gaps are colored gray. The fourth to ninth columns display contigs in the other six assemblies, and in the column, yellow colored contigs are those used for gap closing. The last column shows the number of contigs or corrected reads that close each gap.

Supplemental Table S4: Five large assembly gaps with two recalcitrant ones

The table shows five assembly gaps that we could not close by assembled contigs alone. We attempted to close them using Nanopore reads. Three of them in Chromosome I (nt 4,605,602-4,668,785), Chromosome II (nt 14,525,986-14,566,400), and Chromosome X (nt 4,149,383-4,226,837)

could be spanned by Nanopore reads; however, two gaps in Chromosome III (nt 7,587,315-7,682,900) and Chromosome X (nt 5,215,995-5,284,061) remain to be closed fully. Each gap was surrounded by two different types of tandem repeat expansions. For each tandem repeat, the table shows its unit length, the number of unit occurrences, the repeat length in our VC2010 assembly, the repeat length in the N2 assembly (WormBase release WS220; i.e., UCSC ce10), and the increase in the repeat length from N2 (ce10) to our VC2010. The two remaining recalcitrant gaps are colored gray; in the VC2010 assembly, they have been tentatively filled with 100 N residues.

Supplemental Table S5: Series of connected contigs and error-corrected reads on chromosomes

In the table, along chromosomal positions in the second and third columns, we order primary Canu contigs (colored orange), contigs for closing gaps (yellow), error-corrected reads (light blue), gaps filled by Nanopore reads (blue), two remaining gaps (black), and tandem repeats around the two gaps (gray). The third and second last columns show the starting and ending positions in each contig. The last column displays the number of multiple error-corrected reads spanning each gap to validate the gap closing.

Supplemental Table S6: Effect of polishing the VC2010 genome assembly by Illumina short reads

The table shows the effect of polishing the VC2010 genome assembly by Illumina short reads using Pilon. We quantitatively assessed the effect in terms of mismatch ratios and indel ratios with the N2 reference assembly before and after polishing by Illumina short reads.

Supplemental Table S7: Comparing assembly qualities by comparing rates of mapped VC2010 and CB4856 (Hawaiian) Illumina reads

This table details the origin and mapping rates of Illumina sequence reads from VC2010 and CB4856; neither set of reads was used for our VC2010 assembly, and thus both sets provide independent tests of assembly quality. Reads were mapped to the N2 reference assembly (ce10), our VC2010 assembly, and the VC2010 assembly of Tyson et al. (2018). Although all three assemblies showed high rates of read mapping, our VC2010 assembly consistently exhibited slightly higher rates (96.73% for VC2010 reads, and 96.71% for CB4856 reads) than N2 (96.71% and 96.68%) and noticeably higher rates than Tyson et al.'s VC2010 assembly (94.79% and 94.10%).

Supplemental Table S8: Status of lifted-over N2 genes in the VC2010 assembly

The first subtable (Summary) summarizes the results of attempted liftover from N2 to VC2010, for all canonical genes from WormBase release WS264 (Lee et al. 2018), and with separate classifications for protein-coding genes, ncRNA-coding genes, and pseudogenes. For protein-coding genes, the observed results of liftovers were: failure to lift over at all (no.lift); successful liftover, but with the open reading frame of the gene being broken during liftover (lifted_bad.trans); successful liftover, but with all isoforms changed from their sequence in the N2 reference assembly (lifted_no.ident); successful liftover, with at least one isoform unchanged from N2 but also with at least one isoform changed from N2 (lifted_some.ident); and successful liftover, with all isoforms unchanged in sequence from N2 (lifted_all.ident).

It should be noted that at least one protein-coding gene in N2, *alh-2* (WBGene00000108), was inactivated by a spontaneous deletion in what become the strain VC2010 (M.L.E., unpublished observations). In our analysis here, *alh-2* is classified as 'lifted_bad.trans'; but, in this case, the classification is due to mutation *in vivo* rather than failed liftover *in silico*. In total, there were 148 protein-coding genes in N2 that we failed to lift over to VC2010 (either through complete unmappability, or through untranslatability). It is possible that other loss-of-function mutations than *alh2* exist in VC2010, and are responsible for at least some of these failures to lift over N2 genes. Which other genes have such VC2010-specific inactivating mutations (if any) will require further analysis.

For ncRNA-coding genes and pseudogenes, the observed results of liftovers were: failure to lift over at all (no.lift); successful liftover, but with the coding sequence of the gene being broken during liftover (lifted_bad.coding); successful liftover, but with all isoforms changed from their sequence in N2 (lifted_no.ident); and successful liftover, with all isoforms unchanged in sequence from N2 (lifted_all.ident). For each gene class, the total of canonical genes in the N2 reference assembly is given (total canonical). The next three subtables (Protein-coding, ncRNA, and Pseudogene) list specific outcomes of liftover for each canonical N2 gene, with their status abbreviated as above (e.g., a gene with completely successful liftover and completely unchanged isoforms is annotated as "lifted_all.ident"). For each gene in the protein-coding subtable, the identity (or identities) of any AUGUSTUS gene predictions with overlapping coding nucleotides are also listed.

Supplemental Table S9: Tandem repeats of length >1 kb

The table shows 109 large tandem repeat expansions of >1 kb that have nearly identical units but differ in the number of unit occurrences between the N2 and VC2010 genomes. The table does not list 10 tandem repeat expansions around the five large gaps, because they are shown in Supplemental Table S4. Specifically, for each tandem repeat, the table lists its starting position, ending position, unit length, number of unit occurrences, and unit strings in the N2 (ce10) and VC2010 genomes. To compare individual pairs of tandem repeat occurrences in the two genomes, the 10th column shows the increase (or decrease) in tandem repeat length from N2 to VC2010, and the identity between pairs of unit occurrences. Of 109 pairs, 106 have identical repeat units, but the remaining three have nearly identical repeat units with single mismatches that are colored red in the 11th, 12th, and 13th columns. We checked whether or not each tandem repeat in the VC2010 assembly was covered by raw PacBio reads by inspecting dot plots with alignments of raw PacBio reads (Supplemental Figure S6): the results are given as "Yes", "Partially", or "No", in column N (the second to last column). The last row shows one way of measuring the read coverage for each tandem repeat, namely, the minimum gap length among alignment of reads with the VC2010 assembly.

Supplemental Table S10: Analysis of overlaps between tandem repeat expansions in VC2010 and YAC-derived genomic sequences in N2

The first subtable (Summary) describes the statistical analysis of overlaps. The observed tandem repeat expansions (TREs) comprised 1,257,512 nt in our VC2010 genome assembly. Of 536 YAC sequences that were used to assemble the N2 genome, 494 had coordinates in N2 that could be mapped (lifted over) to coordinates in VC2010, and thus could be checked for overlaps with TREs. These 494 YACs covered 22,649,103 nt (22%) of the VC2010 genome (102,092,263 nt). We found that 511,501 nt of the TREs overlap with the YACs. This was a frequency of 40.7%, 1.8 times higher than the 22% expected randomly ($p < 2.2\times10^{-16}$, two-tailed Fisher test). Therefore, it is not true that YACs account for all TREs; but it is true that TREs arise disproportionately often from N2 regions sequenced from YACs.

The other subtables provide, in GFF3 format (*https://github.com/The-Sequence-Ontology/Specifications/blob/master/gff3.md*), the following data: genomic coordinates in VC2010 of YAC-derived genome sequences in the N2 (WS264) genome assembly (YACs_in_N2.gff3); successfully lifted over from the N2 genome assembly (YACs_lifted_to_VC2010.gff3); genomic coordinates of N2 YACs that could not be lifted to VC2010 (unliftable_YACs_in_N2.gff3); genomic coordinates of tandem repeat expansions in VC2010 (VC2010_TREs.gff3); genomic coordinates of tandem repeat

expansions in VC2010 that overlap with YAC-derived regions of N2 that had been lifted over into VC2010 genomic coordinates (VC2010_TREs.in.YACs.gff3).

Supplemental Table S11: Statistics of repetitive elements of RNA

We examined three known repetitive elements of RNA: 5S RNA, 18S/28S RNA, and pSX1. The third and fourth columns show the respective numbers of occurrences of the three RNA types in the VC2010 assembly and a Nanopore read, indicating an underestimate of RNA occurrences in the VC2010 assembly. The last two columns show the estimated numbers of RNA occurrences that are predicted from Nanopore and PacBio reads.

Supplemental Table S12: Occurrences of 5S RNA, 18S/28S RNA, and pSX1 in the VC2010 assembly

This shows the positions of occurrences of 5S RNA, 18S/28S RNA, and pSX1 in the VC2010 assembly. For each occurrence of 5S RNA, 18S/28S RNA, and pSX1 in the rows, the table also presents the identity ratio with the genome, number of mismatches, number of gaps, E-value of alignment, and bit score.

Supplemental Table S13: Occurrences of 5S RNA, 18S/28S RNA, and pSX1 in Nanopore reads

Similar to Supplemental Table S12, this shows the positions of occurrences of 5S RNA, 18S/28S RNA, and pSX1 in Nanopore reads. The second column shows the read identifiers of Nanopore reads. For each type of RNA (5S RNA, 18S/28S RNA, and pSX1), all occurrences are found within a single Nanopore read that contains all occurrences either of 5S RNA, or of 18S/28S RNA, or of pSX1.

Supplemental Table S14: Telomeres

Comparison of telomeric tandem repeats of unit GCCTAA at the left (5') and right (3') ends of chromosomes between the N2 and VC2010 assemblies. Because the right ends of our initial assemblies for Chromosomes I and III in VC2010 did not exhibit telomeric tandem repeats, we searched for raw Nanopore and PacBio reads that matched the two right end regions, identified such matching Nanopore and PacBio reads, and estimated the respective right end telomeric tandem repeats from those reads, which are colored red in the table. Supplemental Figure S9 also shows how Nanopore and PacBio reads are aligned with the right end regions.

Supplemental Table S15: Statistics of PD2182 and PD2183 genome assemblies

The top table shows the statistics of PacBio subreads collected from PD2182 and PD2183, and the middle table shows the statistics of assembled contigs. In the bottom table, we show the mismatch and insertion/deletion ratios between the N2 reference assembly (ce10) and each of the PD2182 and PD2183 genome assemblies.

Supplemental Table S16: Insertions into the VC2010 assembly

This lists large insertions of size >100 nt into the VC2010 assembly. The 5th, 6th, and 7th columns show the starting, ending positions, and length of an insertion into VC2010, while the second through fourth columns describe the corresponding N2 region (which is not necessarily a single position but is often a range). We here explain the reason. Supplemental Figure S10 shows the dot plots between pairs of corresponding regions in the two genomes. The boundaries of an insertion are often ambiguous and are difficult to determine due to repetitive elements. Some insertions are surrounded by large duplicated regions (*e.g.*, see II:3,882,389-3,885,370 and II:12,748,228-12,748,588 in the VC2010 assembly). Thus, the N2 reference region corresponding to an insertion in the VC2010 assembly is not necessarily a single position but is often represented by a region.

Supplemental Figure S10 also shows dot plots between the VC2010 assembly and each of the assemblies for N2, PD2182, and PD2183; these plots were used to test whether an insertion was a variant in the lineage from N2 to VC2010, or was missing in the N2 reference due to misassembly. When the decision was difficult to make because of large changes in PD2182 and PD2183, we classified it as undetermined. The eight and ninth columns show the respective categorization of decision by using PD2182 and PD2183. The tenth column shows a combined decision of the eight and ninth columns. When one of the categorizations were undetermined but the other was determined, we used the determined decision. When the two decisions were inconsistent, the combined decision was undetermined. Such a case is observed at II:3,882,389-3,885,370 of VC2010. The last three columns summarize the results.

Supplemental Table S17: Deletions from the VC2010 assembly

The table lists deletions from the VC2010 assembly in a manner similar to Supplemental Table S16, and Supplemental Figure S11 shows dot plots corresponding to individual deletions.

Supplemental Table S18: Duplicated regions in the N2 and VC2010 assemblies

In a manner similar to Supplemental Table S16, the table lists duplications detected from the comparison between the N2 and VC2010 assemblies, and Supplemental Figure S12 shows dot plots corresponding to individual duplications in either N2 or VC2010.

Supplemental Table S19: Transposons in the N2 and VC2010 assemblies

The upper table shows the number of occurrences of known transposons in the N2 (ce10) and VC2010 assemblies. Five differences are found, and each is listed in the lower table.

Supplemental Table S20: Status of protein-coding genes predicted with AUGUSTUS in the VC2010 assembly

Each AUGUSTUS gene is listed as either sharing, or not sharing, at least one coding nucleotide with a lifted-over N2 gene. AUGUSTUS genes that share N2 reference coding nucleotides are likely to be alternative predictions of already-known, lifted-over N2 reference genes. AUGUSTUS genes that share no such nucleotides are more likely (though not at all certain) to be novel genes.

Supplemental Table S21: Genes predicted by AUGUSTUS that have some VC2010-specificity

This describes the characteristics of 183 genes predicted by AUGUSTUS that have at least one VC2010-assembly-specific coding nucleotide, and that do not share coding nucleotides with N2 genes lifted over to our VC2010 assembly.

The data columns involving BlastN and BlastP searches have a common purpose: to provide a rapid, heuristic way to make likely inferences about whether a given AUGUSTUS gene prediction is more or less likely to be a genuinely novel one. However, these are not a substitute for manual curation; both 'unlikely' and 'likely' new genes will certainly have counterintuitive results when examined carefully. The first set of BlastN hits, against the N2 reference assembly, allow one to see if an AUGUSTUS gene's coding sequences correspond with exactly identical N2 genomic DNA, or whether there are varying degrees of difference between the AUGUSTUS CDS DNA and N2. In a few striking cases, the similarity to N2 genomic DNA was poor enough that no hit was reported, and these are most likely to be truly new genes. The second set of BlastN hits against VC2010 assembly provide a positive control against which BlastN hits to N2 reference DNA can be compared. BlastP hits to *C. elegans* proteins encoded by N2 genes allow one to see if the predicted AUGUSTUS proteins are identical to N2, or nearly so; BlastN hits to *C. elegans* protein-coding DNA (CDS DNA) provide a more sensitive

test of whether the coding DNA, rather than just the protein, is identical (in principle, one could have ~100% amino acid identity but down to only ~70% nucleotide identity). BlastP searches against *Caenorhabditis nigoni* provide similarities to the proteome of a closely related species with a PacBio-based genome assembly, and thus offer some positive control for genes lost in N2 but detected by PacBio sequencing.

Other forms of gene annotation (e.g., predicting whether the protein product is secreted, and Pfam 31 domains) are standard. For AUGUSTUS genes that seem most likely to be novel, they give partial evidence for their function. Note that similarities to *C. elegans* proteins that are clearly not identities also provide such evidence (by inference from possible paralogy).

The full data columns of this table are as follows:

Gene: a given protein-coding gene predicted by AUGUSTUS 3.3 in the VC2010 assembly. All further data columns are pertinent to that particular gene.

VC2010-spec.: this denotes the degree to which a gene's coding nucleotides fall within VC2010-assembly-specific DNA. Three degrees of specificity are indicated: "1+_nt" (for genes having at least one nucleotide of VC2010-assembly-specific DNA, but less than one full coding exon/CDS), "1+_exons" (for genes having at least one full coding exon, but not all such exons), and "All_exons" (for genes having all of their coding exons in VC2010-assembly-specific DNA).

N2_same.strand: this lists any N2 reference gene that is on the same DNA as the AUGUSTUS gene and that overlaps it, despite the latter putatively having no shared coding nucleotides. In practice, such AUGUSTUS genes are likely to be exons of the spanning N2 gene, particularly given a high BlastP or BlastN score to the N2 gene's protein or CDS DNA.

N2_opposite.strand: this lists any N2 reference gene that is on the opposite DNA as the AUGUSTUS gene and that overlaps it.

N2_genDNA: genomic coordinates for the best-scoring set of nucleotides within the N2 (WS264) reference assembly that was detected by a BlastN search, using the AUGUSTUS gene's longest-isoform CDS DNA as a query.

N2_perc_id: the percentage of nucleotide identity observed for the hit in N2_genDNA.

N2_E-value: the E-value observed for the hit in N2_genDNA.

VC2010_genDNA: genomic coordinates for the best-scoring set of nucleotides within the VC2010 genome assembly that was detected by a BlastN search, using the AUGUSTUS gene's longest-isoform CDS DNA as a query.

VC_perc_id: the percentage of nucleotide identity observed for the hit in VC2010_genDNA.

VC_E-value: the E-value observed for the hit in VC2010_genDNA.

Cel_prot: the best-scoring protein encoded by the N2 reference assembly that was detected by a Blast search, using the AUGUSTUS gene's longest-isoform protein product as a query.

Cel_perc_id: the percentage of amino acid identity observed for the hit in Cel_prot.

Cel_E-value: the E-value observed for the hit in Cel_prot.

Cel_CDS_DNA: the best-scoring set of nucleotides within the coding DNA (CDS DNA) of the N2 reference assembly that was detected by a BlastN search, using the AUGUSTUS gene's longest-isoform CDS DNA as a query.

Cel.CDS_perc_id: the percentage of nucleotide identity observed for the hit in Cel_CDS_DNA.

Cel.CDS_E-value: the E-value observed for the hit in Cel_CDS_DNA.

Cni_prot: the best-scoring protein encoded by the *Caenorhabditis nigoni* genome that was detected by a Blast search, using the AUGUSTUS gene's longest-isoform protein product as a query.

Cni_perc_id: the percentage of amino acid identity observed for the hit in Cni_prot.

Cni_E-value: the E-value observed for the hit in Cni_prot.

Prot_size: this shows the full range of sizes for all protein products from a gene's predicted isoforms.

Max_prot_size: the size of the largest predicted protein product.

Phobius: this denotes predictions of signal and transmembrane sequences made with Phobius 1.01 (Kall et al. 2004). 'SigP' indicates a predicted signal sequence, and 'TM' indicates one or more transmembrane-spanning helices, with N helices indicated with '(Nx)'. Varying predictions from different isoforms are listed.

NCoils: this shows coiled-coil domains, predicted by ncoils (Lupas 1996). Both the proportion of such sequence (ranging from 0.01 to 1.00) and the exact ratio of coiled residues to total residues are given. Proteins with no predicted coiled residues are blank.

Psegs: this shows what fraction of a protein is low-complexity sequence, as detected by pseg (Wootton 1994). As with Ncoils, the relative and absolute fractions of each protein's low-complexity residues are shown.

PFAM: predicted protein domains from Pfam 31 (Finn et al. 2016), with a domain-specific threshold for significance.

Supplemental Table S22: Possible alternative VC2010-assembly-specific exons of N2 reference genes

This describes the characteristics of 13 genes, a subset of the 183 genes in Supplemental Table S21, that were judged likely to represent alternative exons of known N2 genes based on same-strand overlap, BlastP similarity, and BlastN similarity. With the exception of chrV_pilon.g14390 (a likely VC2010-specific exon of the titin ortholog *ttn-1*), these genes have not been manually examined or curated. The data columns here are identical to those in Supplemental Table S21.

Supplemental Table S23: Possible new protein-coding genes in the VC2010 assembly

This table describes the characteristics of 53 genes, a subset of the 183 genes in Supplemental Table S21, that were judged most promising candidates for genuinely new genes, based on less than ~100% identity to N2 genes and on amino acid compositions that were not obviously aberrant. As with all other members of the 183-gene set, manual examination and curation will be required to validate or refute these possible new genes conclusively (although some instances are quite likely to be real, as shown in Figures 4B-C). The data columns here are identical to those in Supplemental Table S21.

Supplemental Table S24: VC2010-assembly-specific RNA-coding loci predicted by INFERNAL/Rfam31

This table gives the genomic coordinates for 29 sites in the VC2010 assembly that fall into VC2010-assembly-specific DNA and that (necessarily) do not overlap N2 genes lifted from N2 to VC2010, along with one or more Rfam 31 motifs identified for each site.

SUPPLEMENTAL DATA FILE LEGENDS

Supplemental Data File S1: Sequence of our final VC2010 genome assembly (version 20180405) in FASTA format.

Supplemental Data File S2: Chain alignment of the N2 reference assembly to our final VC2010 assembly, used for liftover of gene annotations.

Supplemental Data File S3: Regions of the VC2010 assembly that either correspond to the N2 reference by lifting-over, or that cannot be so mapped from N2 (i.e., are VC2010-assembly-specific genomic DNA), in GFF3 format (*https://github.com/The-Sequence-Ontology/Specifications/blob/master/gff3.md***).**

Supplemental Data File S4: Canonical gene annotations lifted over from the N2 reference to the VC2010 assembly, in GTF format (*http://mblab.wustl.edu/GTF2.html***).**

Supplemental Data File S5: Canonical gene annotations that could not be lifted over from the N2 reference to the VC2010 assembly, in GTF format.

Supplemental Data File S6: Protein-coding genes predicted in VC2010 assembly with AUGUSTUS, in GFF3 format.

Supplemental Data File S7: Non-coding RNA genes predicted in the VC2010 assembly with INFERNAL/RFAM, in GFF3 format.

SUPPLEMENTAL METHODS

Creating isogenic sublines

To produce a wild-type reference population of *C. elegans* with maximum isogenicity, we began with VC3540, VC3510 and VC3525, three clonalized derivative strains of VC2010 (Flibotte et al. 2010), which in turn is a derivative strain of N2. Three individual hermaphrodites from VC2010 were picked, and grown clonally in parallel for 10 generations before being established as the strains VC3540, VC3510 and VC3525 (Edgley & Moerman, unpublished). From each of these strains, three single worms were again picked to individual plates to form new subclonal populations. One subclonal population of each, observed with apparently normal wildtype growth, was chosen for further expansion. The subclonal populations were frozen as live strains (Brenner 1974) with the designations PD1073 (derived from VC3540), PD1074 (from VC3510), and PD1075 (from VC3525).

All of our genomic sequence data for the VC2010 genome assembly were generated from these three strains; however, the bulk of these sequence data were generated from PD1074 alone. We have therefore chosen PD1074 as a new reference *C. elegans* laboratory strain, and have provided the *Caenorhabditis* Genetics Center with numerous aliquots to allow long-term genetic reproducibility of this strain in all *C. elegans* laboratories (*https://cgc.umn.edu/strain/PD1074*).

Additional *C. elegans* strains were as follows: PD2182 refers to a population derived from the wild isolate strain CB4856 (originally HA8; often called "Hawaiian"; Hodgkin and Doniach 1997; Thompson et al. 2015), obtained from the CGC in May 2014. PD2183 refers to a population of animals derived from a single cloned animal of the wild isolate strain MY2 (Rockman and Kruglyak 2009).

Genomic DNA extraction of ultra-long DNA using gel plugs

Genomic DNA for PacBio. Genomic DNA was isolated from living and flash-frozen animals approximately as previously described (Shoura et al. 2017) but with the following modifications: lysis was performed at 52°C for 5 hours with no shaking; pipette tips were used to isolate nucleic acid pellets by carefully positioning the tip alongside the pellet and collecting the pellet onto the pipette tip; DNA pellets were air dried for 16 hours following ethanol precipitation; pellets were slowly suspended and dissolved in TE (10 mM TrisCl, 1 mM Na_2EDTA, pH 8.0) at 4°C for at least two hours. Genomic DNA was handled with extra care at all steps and never frozen before sequencing. For long term storage, DNA was stored as ethanol slurries (DNA precipitates under ethanol) at -80°C. The average size of genomic DNA obtained from this method was around 50 kb. DNA obtained from each strain was subjected to Illumina sequencing to verify the strains, assess *E. coli* contamination, and obtain accurate short reads

for assembly corrections. Illumina sequencing libraries were prepared using the Nextera kit according to manufacturer instructions.

Genomic DNA for Nanopore. We extracted maximally high molecular weight (unfragmented) genomic DNA for Nanopore sequencing from agarose plugs approximately as previously described (Schwartz and Cantor 1984) but with the following modifications. Animals were placed live in molten 1.5% low-gelling temperature agarose in TE (10 mM TrisCl, 1 mM Na_2EDTA, pH 8.0) at 50°C. The animal/agarose mixture was pipetted into sterile plastic transfer pipettes (Fisher), where the stem of the pipette served as a mold for the mixture to solidify on ice for around an hour. A sterile razor was used to cut the stem into 1-inch pieces and the gel matrix was released from the mold. Cell lysis was performed in the gel matrix by incubating the resulting cylindrical *C. elegans* gel plugs in 45 ml of worm lysis buffer and proteinase K as described above at 50°C for two days. Gel plugs were then washed in 100 mM TrisCl and 0.5 M EDTA pH 8.0 at least three times, and stored at 4°C. Genomic DNA was purified from the gel plugs using Gelase (Epicentre) according to the manufacturer's recommendations, except for longer incubation times as needed for each gel plug assessed visually by the portion of liquified agarose. The gelase-treated plug was immediately (and very gently) placed on VSWP 0.025-µM membranes (Millipore, Fisher) floating over 10× TE (100 mM Tris–Cl, 10 mM Na_2EDTA, pH 8.0) in a Petri dish. DNA was left to dialyze against 10× TE for 24 hours and 1× TE for the following days. The size of the DNA obtained was checked on pulse-field gels (Schwartz and Cantor 1984) and it was shown to be >1 Mb. Freshly dialyzed genomic DNA was used for the Nanopore library preparation.

Genomic DNA extraction and Illumina sequencing

For the *C. elegans* strains PD1073 (VC3540), PD1074 (VC3510), PD1075 (VC3525), PD2182 (CB4856), and PD2183 (MY2), DNA extraction was carried out approximately as previously described (Sha et al. 2010) except for modifications described here. Worm strains were grown on standard NGM plates prepared with OP50 bacterial lawns. M9 buffer (22 mM KH_2PO_4, 42 mM Na_2HPO_4, 86 mM NaCl, 1 mM $MgSO_4$) was used to collect worms, and wash pelleted worms twice. Worms were then transferred to a 1.5-ml screw-cap tube (Sarstedt), pelleted, and flash frozen in liquid nitrogen. 450 µl Worm Lysis buffer (0.1 M Tris pH 8.5, 0.1 M NaCl, 50 mM EDTA, 1% SDS) plus 20 µl of 20 mg/ml proteinase K in TE pH 7.4, was added to a ~50 µl pellet of frozen worms. The lysis mixture was incubated at 62°C for 45 minutes with periodic vortexing. DNA was precipitated was by adding 180 µl of 6M (saturated) NaCl, centrifuging for 5 minutes at 13,000 g, transferring supernatant to a new tube and repeating centrifugation. Supernatant was transferred to a new 2 ml tube and precipitated with 2

volumes of ethanol and 1 µl of glycogen. After centrifuging at 21,130 g for 15 minutes, the pellet was washed with 100% ethanol and centrifuged again. After drying, the pellet was resuspended in 100 µl TE pH 8.0. RNAse treatment was carried out by adding 1 µl RNAse A from a 1 mg/ml stock, and incubating 1 hour at 37°C. 2 µl of glycogen and 350 µl of 1× STOP buffer (1 M ammonium acetate, 10 mM NaEDTA, 0.2% SDS) was added and mixed by vortexing briefly. DNA was extracted using Phase-lock Heavy tubes (QuantaBio, Beverly Ma, USA), first with 500 µl 1:1 phenol:chloroform, then with 500 µl 100% chloroform. The aqueous layer was transferred to a new 1.5 ml tube and DNA was precipitated by adding 1 ml 100% ethanol. After pelleting DNA and drying the pellet, DNA was resuspended in 100 µl TE buffer (pH 8.0). DNA sequencing libraries were made using a Nextera DNA Sample preparation Kit (Illumina part # FC-212-1031) and were subsequently sequenced on an Illumina MiSeq.

Base calling and error correction of PacBio and Nanopore reads

For handling Nanopore read data, we used two base callers, MinKNOW version 1.7.14/1.10.11/1.10.16 and Albacore version 2.1.7, to process Oxford Nanopore fast5 files. These two basecallers output largely different reads even from an identical DNA fragment (Supplemental Figure S4F), and they have merits and demerits in terms of nucleotide accuracy and noise such as large insertions of random bases. Thus, we used Nanopore reads from both basecallers so as to increase the possibility of linking PacBio contigs by Nanopore reads. We also corrected errors in Nanopore reads using the correction and trimming steps of Canu version 1.3; however, we found that reads were likely to be longer before error-correction. Thus, we used raw Nanopore reads before error-correction for linking PacBio contigs, in addition to error-corrected reads.

Genome assembly

For genome assembly, we used two datasets of raw reads; one set had only PacBio reads for generating contigs with highly accurate bases, while the other had both PacBio and Nanopore reads for generating longer contigs that could close gaps. We applied Canu, FALCON, and miniasm to both datasets, while applying HINGE to the PacBio/Nanopore dataset alone. This resulted in seven different initial VC2010 genome assemblies (Supplemental Table S2).

Closing gaps with multiple assemblies

We connected two contigs when they had identical subregions of ≥100 kb at their ends. When a gap was spanned by single contigs in two or more genome assemblies, we selected contigs assembled from PacBio reads because their nucleotide accuracy was expected to be higher than that of contigs from both PacBio and Nanopore reads.

Checking proximity of contig pairs around five large gaps

We first identified pairs of contigs around five large gaps by aligning assembled contigs onto the N2 reference genome. To confirm the proximity between these pairs from an independent source, we used Hi-C reads collected from the same strain, PD1074 (accession numbers: SRR3105476, SRR3105477), and aligned pairs of Hi-C reads to the VC2010 genome using Juicer (version 1.5.6; *https://github.com/aidenlab/juicer/releases/tag/1.5.6*; Durand et al. 2016). We divided the entire VC2010 genome into 5-kb non-overlapping bins, calculated the average number of aligned read pairs between two bins at distance x (multiples of 5-kb), and determined the average frequency distribution of Hi-C read pairs that linked two arbitrary bins at distance x. We used 40 values for x, ranging from 0 to 195 kb. The average number of Hi-C read pairs serves as a good indicator for estimating the size of a gap, because a higher average number indicates a smaller gap size.

Subsequently, for five gaps in Supplemental Table S4 (I:4,605,602-4,668,785; II:14,525,986-14,566,400; III:7,587,315-7,682,900; X:4,149,383-4,226,837; and X:5,215,995-5,284,061), we analyzed their repeat element copy numbers and estimated their respective sizes as 65 kb, 50 kb, >100 kb, 80 kb, and >75 kb. Using the average frequency distribution of Hi-C read pairs at distance x, the average numbers of Hi-C read pairs for the five gaps were predicted to be 4.5, 5.3, 3.7, 4.1, and 4.2; the observed numbers were 4, 9, 7, 3, and 3. These numbers support the proximity of contig pairs around individual gaps, which we had inferred through synteny with the N2 assembly.

Filling five large gaps with Nanopore long reads

A tandem repeat (expansion) is a repetition of a string pattern of one or more nucleotides. When the pattern is minimal in the sense that it is not a tandem repeat of a smaller pattern, it is called a unit. For example, CA is the unit in CACACACA, but CACA is not because it is a tandem repeat of CA. In reality, the definition needs to be somewhat nuanced so that a tandem repeat can be a repetition of a unit or its slight variant. To determine the unit string of each tandem repeat, we used contigs based solely on PacBio data, because their nucleotide accuracy was expected to be higher than contigs based on both

PacBio and Nanopore data (Supplemental Figure S4). Each of the five large gaps was spanned by multiple Nanopore reads. To determine the length of each tandem repeat from those gap-spanning Nanopore reads, we selected the longest tandem repeat that matched tandem copies of the unit with the best or second-best nucleotide accuracy, and used its length for the tandem repeat. This method allowed us to refine tandem repeats around the gaps in our genome assembly (Supplemental Table S4); these tandem repeats are underestimated or erroneous in the N2 reference assembly (Supplemental Figure S5).

Tandem repeats around gaps were underestimated in assembled contigs; hence these needed to be determined accurately, with the task being non-trivial. When multiple Nanopore reads covered a tandem repeat, the length of the tandem repeat were frequently inconsistent among the multiple reads, and long noisy sequences were often inserted into reads at random positions (Supplemental Figure S4B). To resolve this discrepancy, we first compared two different basecallers, MinKNOW and Albacore; however, these two basecallers often called different sequences for an identical DNA fragment, which in turn made the lengths of focal tandem repeats different (Supplemental Figure S4F). For instance, one basecaller could remove long noisy sequences but produce shorter reads that failed to span a gap. Given the trade-off between error-correction and length in Nanopore reads, we searched reads both before and after error-correction for those that could bridge gaps.

Statistical analysis of different mapping rates for Illumina MiSeq reads mapped to the three genome assemblies

Supplemental Table S7 compares the numbers of Illumina MiSeq reads that were aligned with the three genome assembly. The numbers of our VC2010 assembly were larger than those of the other two; we thus asked whether the differences were significant by testing the null hypothesis that reads were mapped independently to both the N2 reference assembly (ce10) and our VC2010 assembly. Because the number of mapped reads was much larger than the number of unmapped reads in each assembly (see the 2×2 contingency tables in Supplemental Table S7), we used the two-tailed Fisher's exact test. We found that the null hypothesis was rejected for all pairs of the three genome assemblies in each of the two read datasets, because *p*-values were smaller than 2.2×10^{-16}. We used the fisher.test function in R.

Assessing coverage of BUSCO genes in genome assemblies

BUSCO aligns a set of taxon-conserved reference genes with a genome assembly and categorizes alignments into one of complete, fragmented, and missing classes; this provides a

measurement of the assembly's completeness and quality (Waterhouse et al. 2018). In our hands (Table 1), BUSCO reported that alignments of four genes (*rpn-9, rps-10, ptps-1,* and D2030.4) were fragmented and three genes (*mog-4*, *scpl-1*, and *rsd-3*) were missing in both the N2 assembly (ce10) and our VC2010 genome assembly; however, we could align all of the seven genes to each of the two assemblies using minimap2. Specifically, the alignments of four genes were fragmented on one strand of the VC2010 assembly but were complete on the reverse strand. With regard to the three missing genes, when there are more than one candidate region where an input gene is mapped to, BUSCO selects one with the lowest E-value score of tblastn and aligns the focal gene with the selected region only, thereby overlooking a complete alignment. A typical example was *mog-4*. Although *mog-4* had a complete alignment in II:14547946-14564322 of N2 (ce10), the complete alignment was overlooked, and *mog-4* was partially aligned with another locus by BUSCO. To settle the problem, we instead aligned each of the three genes with all candidate regions, selected the best alignments, and checked them manually.

E. coli genome sequences used to analyze Illumina reads unmapped to the VC2010 genome assembly

We used genome sequences for two *E. coli* strains, K-12 substrain MG1655 (accession NC_000913) and C chromosome (accession CP029371) and mapped previously unmappable Illumina reads from VC2010 to these *E. coli* genomes (Supplemental Table S7) using bwa mem (ver.0.7.16a; Li and Durbin 2009) with default parameter settings.

Determining genomic overlaps between alternatively predicted genes

There are two senses in which two different predicted genes may overlap with one another. They may share a common span of nucleotides in genomic DNA, and this span may be defined either by their complete extent (5' end to 3' end, including introns) or their exonic boundaries alone. Alternatively, their overlap may be more strictly defined by requiring not merely that they share one or more nucleotides in the genome, but also that these nucleotides should be used for coding in identical ways (i.e., for both genes, the shared nucleotide should be on the same DNA strand and in the same reading frame). We used the latter definition (exactly overlapping coding) to identify which new predictions of protein-coding genes with AUGUSTUS overlapped lifted-over genes from N2, to avoid mistakenly equating genes whose products were not actually equivalent. We used the former, broader definition (shared genomic coordinates) to check for possible instances where a new "gene" was actually a new exon inside an older lifted-over N2 gene. We also (necessarily) used the broader definition of overlap when

checking new gene predictions for overlap with novel, VC2010-assembly-specific regions (i.e., regions of the VC2010 assembly that had no corresponding sequences in N2 that could be defined by chain alignment and annotation liftover). 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*). This script considers only nucleotides of protein-coding exons (defined in GFF3 annotations with 'CDS'). To carry out all other tests of overlap (most importantly, to identify which of our gene predictions fell completely within VC2010-assembly-specific genome regions) we used the intersect program of BEDTools 2.27.1 to identify which genes fell completely within VC2010-assembly-specific genome regions. For protein-coding genes, we tested both whether the entire span of a gene fell in VC2010-assembly-specific blocks, and whether all of their CDS exons fell into them (considering the possibility that an intron might contain N2 DNA). In practice, all VC2010-assembly-specific protein-coding genes could be identified simply by considering their gene spans. Likewise, we used BEDTools to test our ncRNA gene predictions for being completely within VC2010-assembly-specific DNA.

Determining and statistically analyzing genomic overlaps between tandem repeat expansions (TREs) in VC2010 and YAC-derived genome sequences in N2

To determine which nucleotide coordinates in the N2 WS264 genome assembly were derived from cloning and sequencing of YACs, we first downloaded and decompressed the GFF3-formatted genome annotations for N2 from WormBase WS264 (*ftp://ftp.wormbase.org/pub/wormbase/releases/WS264/species/c_elegans/PRJNA13758/c_elegans.PRJNA13758.WS264.annotations.gff3.gz*). We extracted a subset of canonical genomic clones with the command "cat c_elegans.PRJNA13758.WS264.annotations.gff3 | grep Genomic_canonical | grep assembly_component > Genomic_canonical.assembly_component.gff3". This subset annotation file gave exact coordinates for clones from which N2 had been assembled, but did not specify which of these clones were YACs. To determine this, we downloaded the full database for WormBase WS264 (*ftp://ftp.wormbase.org/pub/wormbase/releases/WS264/acedb*), decompressed it with its INSTALL shell script, and extracted data from it with the TableMaker function of xace from the ACeDB software suite. In order to run ACeDB, we downloaded its actively maintained NCBI software fork (*ftp://ftp.ncbi.nlm.nih.gov/repository/acedb/Software/acedb/acedb.source.tar.gz*), and compiled it with the commands "cd [source code directory]; export ACEDB_MACHINE=ICC_centos7 ; make ;". With TableMaker, we extracted hits to the query "Clone => Type => YAC" as tab-delimited data; with these data and Genomic_canonical.assembly_component.gff3, we

used Perl to extract GFF3-annotated names and genomic coordinates for 536 YAC clones used to sequence *C. elegans* N2 (Supplemental Table S10, YACs_in_N2.gff3). As we had previously done for N2 gene annotations, we used UCSC liftOver and an N2-VC2010 chain alignment to map these YAC annotations from N2 to VC2010 genomic coordinates. This yielded 494 YACs that were mapped to VC2010 (Supplemental Table S10, YACs_lifted_to_VC2010.gff3) and 42 that could not be (Supplemental Table S10, unliftable_YACs_in_N2.gff3). We used the getfasta function of BEDTools to extract the subset of VC2010 genomic sequences corresponding to N2 YACs, and used count_fasta_residues.pl (available at *https://github.com/SchwarzEM/ems_perl/tree/master/fasta*) to determine its total size (22,649,103 nt, out of 102,092,063 nt; i.e., YACs mapped to VC2010 comprise 22.2% of the whole genome).

Nucleotide coordinates for 114 TREs in VC2010 were extracted from Supplemental Table S5 and Supplemental Table S9, and reformatted into GFF3 format with Perl (Supplemental Table S10, VC2010_TREs.gff3). Their total size in the VC2010 genome was determined with BEDTools/getfasta and count_fasta_residues.pl (114 sequences, totalling 1,257,512 nt).

The overlap of TREs and N2 YACs mapped to VC2010 was determined with the intersect function of BEDTools (Supplemental Table S10, VC2010_TREs.in.YACs.gff3); the number and total size of the overlap was determined with BEDTools/getfasta and count_fasta_residues.pl (56 sequences, totalling 511,501 nt; this corresponded to 49.1% of all TRE sequences and 40.7% of all TRE nucleotides. These proportions are 2.2-fold and 1.8-fold greater than the 22.2% expected by chance.

Statistical significance of these disproportionate overlaps was determined by the two-tailed exact Fisher test, as implemented in the binom.test function of R, with a 99% confidence interval.

SUPPLEMENTAL DATA ACCESS AND STRAIN AVAILABILITY

The genome assembly described here, VC2010-1.0, has been deposited at the ENA (study accession, PRJEB28388; assembly accession, GCA_900538205) and has been imported to WormBase (releases WS268 onward). Sequence read data have been deposited at the NCBI SRA: PacBio, Nanopore, and Illumina read sequence data for VC2010 (BioProject accession: PRJNA430756); PacBio and Illumina data for PD2182 (BioProject accession: PRJNA482888); and PacBio data for PD2183 (BioProject accession: PRJNA482889). Seven gapped VC2010 genome assemblies that were used as the precursors of our final VC2010 genome assembly, have been deposited in a permanent data archive at the Open Science Framework (*https://osf.io/jx89y*; doi:10.17605/osf.io/jx89y). A contaminant-free (*C. elegans*-only contigs) version of the VC2010 genome assembly by Tyson et al. has also been

deposited (with the permission of Tyson et al.) in a permanent data archive at the Open Science Framework (*https://osf.io/dbgkm*; doi:10.17605/osf.io/dbgkm).

The *C. elegans* strain 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 VC2010-1.0 genome assembly.

REFERENCES

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

Durand NC, Shamim MS, Machol I, Rao SS, Huntley MH, Lander ES, Aiden EL. 2016. Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Syst* **3**: 95-98.

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**: D279-285.

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.

Hodgkin J, Doniach T. 1997. Natural variation and copulatory plug formation in *Caenorhabditis elegans*. *Genetics* **146**: 149-164.

Kall L, Krogh A, Sonnhammer EL. 2004. A combined transmembrane topology and signal peptide prediction method. *J Mol Biol* **338**: 1027-1036.

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.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**: 1754-1760.

Lupas A. 1996. Prediction and analysis of coiled-coil structures. *Methods Enzymol* **266**: 513-525.

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

Schwartz DC, Cantor CR. 1984. Separation of yeast chromosome-sized DNAs by pulsed field gradient gel electrophoresis. *Cell* **37**: 67-75.

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.

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.

Thompson OA, Snoek LB, Nijveen H, Sterken MG, Volkers RJ, Brenchley R, Van't Hof A, Bevers RP, Cossins AR, Yanai I et al. 2015. Remarkably divergent regions punctuate the genome assembly of the *Caenorhabditis elegans* Hawaiian strain CB4856. *Genetics* **200**: 975-989.

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.

Waterhouse RM, Seppey M, Simao 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.

Wootton JC. 1994. Non-globular domains in protein sequences: automated segmentation using complexity measures. *Comput Chem* **18**: 269-285.