HIGHLIGHTED ARTICLE
GENETICS | INVESTIGATION

Remarkably Divergent Regions Punctuate the Genome Assembly of the *Caenorhabditis elegans* Hawaiian Strain CB4856

Owen A. Thompson,* L. Basten Snoek,† Harm Nijveen,‡ Mark G. Sterken,† Rita J. M. Volkers,† Rachel Brenchley,§ Arjen van't Hof,§ Roel P. J. Bevers, Andrew R. Cossins,§ Itai Yanai,†† Alex Hajnal,‡‡ Tobias Schmid,‡‡ Jaryn D. Perkins,§§ David Spencer,* Leonid Kruglyak,*** Erik C. Andersen,††† Donald G. Moerman,§§ LaDeana W. Hillier,* Jan E. Kammenga,† and Robert H. Waterston*,[1]**

*Department of Genome Sciences, University of Washington, Seattle, Washington 98195, †Laboratory of Nematology, Wageningen University, 6708 PB Wageningen, The Netherlands, ‡Laboratory of Bioinformatics, Wageningen University, NL-6708 PB Wageningen, The Netherlands, §Centre for Genome Research, Institute of Integrative Biology, University of Liverpool, Liverpool L69 7ZB, United Kingdom, **Laboratory of Systems Biology and Genetics, Institute of Bioengineering, School of Life Sciences, Ecole Polytechnique Fédérale de Lausanne, CH-1015 Lausanne, Switzerland, ††Department of Biology, Technion-Israel Institute of Technology, Haifa 32000, Israel, ‡‡Institute of Molecular Life Sciences, University of Zurich, CH-8057 Zurich, Switzerland, §§Department of Zoology and Michael Smith Laboratories, University of British Columbia, Vancouver, BC, Canada V6T 1Z3, ***Howard Hughes Medical Institute, Department of Human Genetics and Department of Biological Chemistry, David Geffen School of Medicine, University of California, Los Angeles, California 90095, and †††Department of Molecular Biosciences, Northwestern University, Evanston, Illinois 60208

ABSTRACT The Hawaiian strain (CB4856) of *Caenorhabditis elegans* is one of the most divergent from the canonical laboratory strain N2 and has been widely used in developmental, population, and evolutionary studies. To enhance the utility of the strain, we have generated a draft sequence of the CB4856 genome, exploiting a variety of resources and strategies. When compared against the N2 reference, the CB4856 genome has 327,050 single nucleotide variants (SNVs) and 79,529 insertion–deletion events that result in a total of 3.3 Mb of N2 sequence missing from CB4856 and 1.4 Mb of sequence present in CB4856 but not present in N2. As previously reported, the density of SNVs varies along the chromosomes, with the arms of chromosomes showing greater average variation than the centers. In addition, we find 61 regions totaling 2.8 Mb, distributed across all six chromosomes, which have a greatly elevated SNV density, ranging from 2 to 16% SNVs. A survey of other wild isolates show that the two alternative haplotypes for each region are widely distributed, suggesting they have been maintained by balancing selection over long evolutionary times. These divergent regions contain an abundance of genes from large rapidly evolving families encoding F-box, MATH, BATH, seven-transmembrane G-coupled receptors, and nuclear hormone receptors, suggesting that they provide selective advantages in natural environments. The draft sequence makes available a comprehensive catalog of sequence differences between the CB4856 and N2 strains that will facilitate the molecular dissection of their phenotypic differences. Our work also emphasizes the importance of going beyond simple alignment of reads to a reference genome when assessing differences between genomes.

KEYWORDS *C. elegans*; genome assembly; evolution; variation

DNA sequence variation, whether present in natural populations or induced in the laboratory, has been central to the functional understanding of genes and genomes. Natural variation has proven particularly valuable in the analysis of quantitative traits while also providing insights into the evolutionary processes that shape genomes. At the same time, mutations of strong phenotypic effect have long been a pillar of experimental genetics. As rapidly improving DNA sequencing technology has simplified both the detection and the cataloging of variation, major efforts have been undertaken to describe variation and then analyze quantitative traits in wild isolates of various model organisms, including *Caenorhabditis elegans*, *Saccharomyces cerevisiae*, *Drosophila*, and *Arabidopsis* (Schacherer *et al.* 2009; Cao

doi: 10.1534/genetics.115.175950
Manuscript received March 2, 2015; accepted for publication April 29, 2015; published Early Online May 19, 2015.
Available freely online through the author-supported open access option.
Supporting information is available online at www.genetics.org/lookup/suppl/doi:10.1534/genetics.115.175950/-/DC1.
Sequence data from this article have been deposited in GenBank under accession no. JZEW00000000 (CB4856 genomic assembly) and in Sequence Read Archive under accession no. SRX1001806.
[1]Corresponding author: Department of Genome Sciences, University of Washington School of Medicine, 3720 15th Ave. NE, Box 355065, Seattle, WA 98195-5065. E-mail: watersto@u.washington.edu

et al. 2011; Andersen *et al.* 2012; Mackay *et al.* 2012; http://www.1001genomes.org).

For *C. elegans*, the canonical wild-type laboratory strain N2 was derived from an isolate found in 1951 in mushroom compost in England (Nicholas *et al.* 1959) and maintained in liquid culture on agar slants and then on *Escherichia coli* until protocols were developed in 1969 that allowed storage of frozen stocks (Sulston and Brenner 1974; Sterken *et al.* 2015). It was the first multicellular organism to have a fully sequenced genome (*C. elegans* Sequencing Consortium 1998), and this sequence has served as the reference for *C. elegans*. Prior studies that have utilized the genetic diversity available in wild populations of *C. elegans*, whether studying specific phenotypes, genes, gene classes, or population mutational spectra, have reported their results with respect to the N2 strain.

Among the many wild isolates of *C. elegans*, one of the most genetically divergent and most heavily studied is CB4856, which was isolated in 1972 by Linda Holden from a pineapple field on the Hawaiian island of Maui (under the name HA8) (Hodgkin and Doniach 1997). It shows multiple phenotypic differences with N2, including production of a copulatory plug, aggregation behavior, a lack of temperature-size dependence, growth rate, fecundity, RNA interference insensitivity by feeding and drug resistance (de Bono and Bargmann 1998; Kammenga *et al.* 2007; Ghosh *et al.* 2012; Pollard and Rockman 2013; Andersen *et al.* 2014), and gene expression differences (Capra *et al.* 2008; Rockman *et al.* 2010; Vinuela *et al.* 2012; Volkers *et al.* 2013). Various populations of recombinant inbred lines (RILs) and a population of introgression lines (ILs) have been generated between CB4856 and N2 to define the genetic architectures of complex genetic traits (Li *et al.* 2006; Rockman and Kruglyak 2008; Doroszuk *et al.* 2009; Andersen *et al.* 2015). Molecular genetic analyses of the Hawaiian strain have revealed polymorphisms associated with several of the above traits as well as others. An online database, WormQTL, has been created for the deposition of expression quantitative trait loci (Snoek *et al.* 2013, 2014; van der Velde *et al.* 2014).

The elucidation of sequence variants in CB4856 has occurred in several steps. Initially, random genomic fragments were compared to the N2 reference genome, revealing >6000 SNVs and small insertion/deletions (indels) (Wicks *et al.* 2001). A later study increased the number of SNVs to >17,000 (Swan *et al.* 2002). The genomic positions of these SNVs are distributed nonrandomly, with more variation present on chromosome arms than in the centers where recombination is lower (Koch *et al.* 2000; Wicks *et al.* 2001). These variants provided suitable markers for genetic mapping using a variety of methods. D. Spencer and R. H. Waterston (unpublished results) cataloged >100,000 SNVs using an early version of massively parallel sequencing (MPS) technology in a whole-genome shotgun (WGS) approach and deposited these variants in WormBase, noting multiple ~25- to 100-kb regions of poor read alignment, possibly due to high sequence divergence. These regions were most prevalent on the left arms of chromosomes I and II along with both arms of chromosome V. Array comparative hybridization identified large copy number variations (CNVs) and found that these CNVs also were enriched on chromosome arms, affecting primarily gene family members that had undergone recent expansion in *C. elegans* (Maydan *et al.* 2007, 2010). A study of chemoreceptor gene families uncovered functional genes in CB4856 that are defective in N2 (Stewart *et al.* 2005). Recent genomic analyses of CB4856 and N2, alongside other isolates, again found the Hawaiian strain to be among the most divergent, either by using sequencing restriction-site-associated DNA markers in 202 strains (Andersen *et al.* 2012) and/or by comparing hybridization of coding sequences between N2, CB4856, and a panel of 46 wild isolates (Volkers *et al.* 2013). Recently, we used MPS to obtain deep WGS coverage, providing a more complete list of differences including indels of a full range of sizes between the N2 reference and the Hawaiian genome (175,097 SNVs and 46,544 indels) (Thompson *et al.* 2013). Another group extended the set further using deeper WGS coverage along with longer reads from the 454 platform (Vergara *et al.* 2014).

One shortcoming of all of these studies has been that they have relied on alignment of the sequence reads to the N2 reference genome. As a result, multiple regions of the Hawaiian genome remain missing or poorly defined. These missing regions include insertions in the Hawaiian genome relative to N2. But, in addition, inspection of the deep WGS coverage revealed some regions of the genome that apparently were so divergent that aligned reads were sparse to absent (D. Spencer, O. A. Thompson, and R. H. Waterston, unpublished results). The sequence of these highly divergent regions and Hawaiian specific sequences must be determined to interpret more fully any genotypic and phenotypic differences between the Hawaiian and N2 strains.

Accordingly, we have undertaken the construction of a Hawaiian reference genome sequence that more completely reflects the sequence differences between the two isolates. To accomplish this goal, we took advantage of several very deep coverage MPS data sets for the Hawaiian genome, a *de novo* assembly program (Chu *et al.* 2013), end sequences from a fosmid library for the Hawaiian genome, recently released RNA-seq data, and low-coverage genome sequence data from 49 RILs (Li *et al.* 2006) and 60 ILs (Doroszuk *et al.* 2009) (Table 1). Exploiting these resources and using a variety of software tools, we have modified the N2 reference genome to generate a draft reference sequence for the Hawaiian genome. The results reveal >60 regions with haplotypes that are substantially divergent from N2. The distribution of these haplotypes in other wild isolates suggests that these regions were present in the genomes of ancestral populations before the world-wide distribution of the *C. elegans* species and have been maintained since that time.

Table 1 CB4856 sequence resources

Data set	PI	Type	Platform	[SIP]E, length (bp)	Insert size	Clones/total bases in reads	Coverage expected (%)
Princeton University	Andersen	DNA	Illumina	PE 104, 104	321 bp	34,711,778/7,220,049,824	69.52× (96.6)
University of Washington (Thompson *et al.* 2013)	Waterston	DNA	Illumina	PE 76, 76	179 bp	21,252,827/3,230,429,704	31.10× (96.6)
Technion	Yanai	DNA	Illumina	PE 100, 100	221 bp	79,406,930/15,881,386,000	80.08× (50.6)
University of Zurich	Hajnal	DNA	Illumina	PE 101,101	484 bp	825,754/166,799,884	1.41× (84.9)
University of Zurich	Hajnal	DNA	SOLiD	PE 50, 35	124 bp	15,760,405/2,679,268,850	7.20× (26.9)
University of British Columbia (Perkins 2010)	Moerman	DNA	Sanger	PE ~770 bp	~33 kbp	15,360/20,520,434	0.20× (97.2)
Washington University (Wicks *et al.* 2001)	Waterston	DNA	Sanger	SE ~764 bp	NA	11,541/8,843,526	0.07× (81.7)
Wageningen University/ University of Liverpool	Kammenga/ Cossins	DNA (ILs/RILs)	SOLiD	SE 50	NA	2,709,932,329/135,496,616,450	766.85× (56.8)
Total DNA							956.43×

SE = single end
PE = paired end

Materials and Methods

Sequencing methods

DNA from the *C. elegans* CB4856 strain was extracted using the Qiagen Blood and Tissue kit and quantified using a Qubit 2.0 broad-range kit. DNA was sheared in a Covaris LE220 sonicator to a size of 300–600 bp and then 400- to 600-bp fragments were gel-extracted after standard Illumina TruSeq sample preparation. One Illumina HiSeq2000 lane was run to obtain the 101-bp paired-end sequence reads used in this study.

Creating a Hawaiian reference

Using a strategy similar to that employed in the analysis of different *Arabidopsis* accessions (Gan *et al.* 2011; Schneeberger *et al.* 2011), we first aligned the random genomic reads to the N2 reference genome, identified SNVs and indels, modified the N2 reference accordingly, and realigned the reads, repeating the process 19 times to create a first version of the Hawaiian genome (20 cycles total) (Figure 1; Supporting Information, Figure S1; Table S1). This process allowed extension of sequence into regions of high divergence, closed large deletions, and built sequence into insertions (Figure 1 and Figure 2). We used the JR-Assembler (Chu *et al.* 2013) to create *de novo* assemblies of the same sequence reads, assessed their quality using the program REAPR (Hunt *et al.* 2013), breaking contigs as needed, and aligned the resultant contigs to the Hawaiian genome. To identify deletions previously missed, we scanned the genome for regions devoid of coverage, merging adjacent regions if they were separated only by short segments of either very low coverage or repeated sequences. For regions flanked by adjacent segments of the *de novo* assembled contigs, we used the contig to close the gap. To confirm that such segments were properly placed in the genome, we used the RIL data to establish their chromosomal location (Figure S2; Figure S3; Table S2). Specific methods for each of these steps are presented in the Supporting Information.

The result is an initial draft reference Hawaiian genome with a total length of 98.2 Mb. Regions of excess coverage (>99×) suggest that we have failed to represent some duplicated segments, which total some 0.5 Mb in length. Also, the *de novo* assembly generated 22 contigs of 16 kb total length that we were unable to locate in the reference. Just as the N2 reference has been improved through continuous community input, we expect users will provide improvements here.

ALE scoring of divergent regions in wild isolates

MPS sequence reads from each of the 39 wild isolates previously studied (Thompson *et al.* 2013) were aligned against both the N2 and CB4856 reference sequences. The resultant alignments were scored using the ALE program (Clark *et al.* 2013). For each divergent region (Table S3), we then plotted the placement score for each strain against the two genomes. Many sites followed a simple binary pattern, with scores for each strain against N2 and CB4856 resembling one of the controls (Figure S4A). Other regions in some strains showed intermediate scores against either or both N2 and CB4856. Inspection of the alignments and ALE score patterns across the region suggested that the strain had intermediate divergence across the region, where some blocks of a region resembled the N2 haplotype and others resembled CB4856 (Figure S4B). These patterns were consistent with the idea that recombination had occurred within the region. However, some regions had one or more strains with reads that aligned poorly with both strains, and inspection of those regions in these strains was consistent with the presence of a third haplotype for the region (Figure S4C). Other regions, particularly those from the left arm of chromosome II, had more complex patterns and were not analyzed further.

Comparing the two reference genomes

The *C. elegans* N2 genome, version WS230 (note that while annotations changed, the N2 genome sequence remained

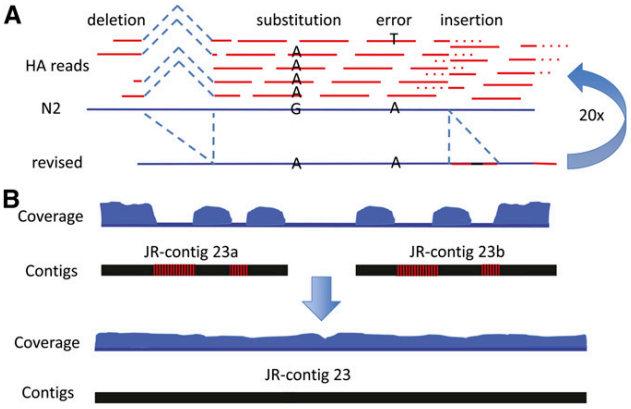

Figure 1 Strategy for constructing a Hawaiian reference sequence. (A) Alignment of 100-bp paired-end reads from the CB4856 genome to the N2 genome. Sites that differed by base substitution and insertion and deletion were recognized, and the N2 genome was altered at those sites. For insertions larger than a read and at the edge of divergent regions, the consensus sequences from the unmatched segments of the reads were added to the reference. Then the reads were aligned to the modified reference, and the cycle was repeated for 20 times, by which time few changes were being made. (B) After the 20 cycles of alterations, areas with incomplete coverage still persisted. To correct these areas, individual reads were assembled *de novo* with the JR-Assembler and aligned against the modified reference. Typically, these JR contigs would show good agreement where read coverage was good, and thus corrections had been made, but poor alignment where the reference sequence did not have coverage and had not been altered from the N2 reference. The JR contigs were also aligned against sequence reads from RILs and ILs. Only RILs and ILs containing a segment of the Hawaiian genome that spanned the JR contig yielded good coverage across these divergent regions, thereby locating the JR contigs on the genome. Where the JR contigs had regions of good match against the reference and their location was confirmed by alignment of reads from RILs and ILs, they were spliced cleanly into the reference. Remaining large deletions were also removed.

identical from WS215 through WS234), was aligned against the Hawaiian genome using LASTZ [version 1.02.00; O = 400, E = 30, K = 3000, L = 3000, M = 0; (Harris 2007)]. The program LASTZ, like BLASTZ (Schwartz *et al.* 2003), uses a series of steps consisting of seeding (a user-specific seed pattern is allowed), gap-free extension, chaining, anchoring, gapped extension, and interpolation. However, unlike BLASTZ, LASTZ can derive its own suitable tuning parameters from the sequences themselves. We chose LASTZ because it is able to perform pairwise gapped full chromosome-to-chromosome alignments using very little memory.

The LASTZ alignments were performed with the chain option and lav output format. Files in the lav output format were converted to psl format using lav2psl (J. Kent, http://hgdownload.cse.ucsc.edu/admin/jksrc.zip) and were chained [creating a sequence of gapless aligned blocks with no overlapping blocks (Kent *et al.* 2003)] using axtChain (-linearGap = loose). The resulting alignment files (a readable text version for coordinate lookup is available in File S1) were merged using chainMergeSort and prepared for the netting step using chainPreNet and then netted (a net is a hierarchical collection of chains, with the highest-scoring nonoverlapping chains on top; their gaps were filled in where possible by lower-scoring chains; for more information see http://genomewiki.ucsc.edu/index.php/Chains_Nets), retaining a unique, ordered set of alignments within the N2 genome using chainNet. The resulting output contains a single set of ordered alignments in the N2 genome followed by additional "fill in" alignments. Only alignments to the same chromosome in N2 and Hawaiian and on the same strand were retained. We reviewed the additional "fill in" alignments and found only a possible 2 kb of unique additional aligned sequence that could have been used for SNV identification. Therefore, we chose not to retain the "fill in" alignments (they occur in the output file after the initial list of the entire N2 chromosomal alignments). The axt files created by the final netting step were parsed using a custom script to create files listing the SNVs and insertion/deletion differences.

When there were gaps in either the N2 or Hawaiian genome between the "major" alignments, they were annotated as indels. Two types of regions were identified between neighboring blocks: (1) simple indels, where the sequence was cleanly inserted/deleted in one genome relative to the other and (2) 816 cases where sequence was present in both the N2 and Hawaiian genomes but could not be accurately aligned (alignment using various Smith–Waterman algorithm implementations revealed different alignments in each case). Thus, we elected to simply retain those regions as blocks of indels. To identify high-quality substitutions within the LASTZ alignments, we required at least three reads and ≤150 reads with the fraction of reads that disagree with the total reads at the site as <0.2.

RNA-seq ALIGNMENTS/TOPHAT/CUFFLINKS

Hawaiian reads from available RNA-seq projects (313,124,440 reads) (Stoeckius *et al.* 2014) were aligned to the Hawaiian genome using TopHat (–b2-fast) (Trapnell *et al.* 2012), and BAM files were generated containing the resulting 215,316,914 aligned reads. Cufflinks (–min-frags-per-transfrag 5–min-intron-length 25–trim-3-avgcov-thresh 5 -p 8) created 31,965 transcript predictions (30,199 genes) composed of 91,185 exons (84,461 different exons) (Trapnell *et al.* 2012).

Genefinder

Using the *C. elegans* table files, we ran Genefinder (version 1.1; P. Green, unpublished results; -orfcutoff -0.5 -intron3-cutoff -0.5 -intron5cutoff -0.5) to provide *de novo* gene prediction in the Hawaiian genome. A total of 22,261 transcripts were predicted, which were composed of 126,518 different exons.

Identifying N2 genes in Hawaiian

To identify *C. elegans* N2 genes in Hawaiian, the N2 predicted genes (WS240) were aligned against the Hawaiian genome using blat (-q = rna). Each N2 gene was aligned to

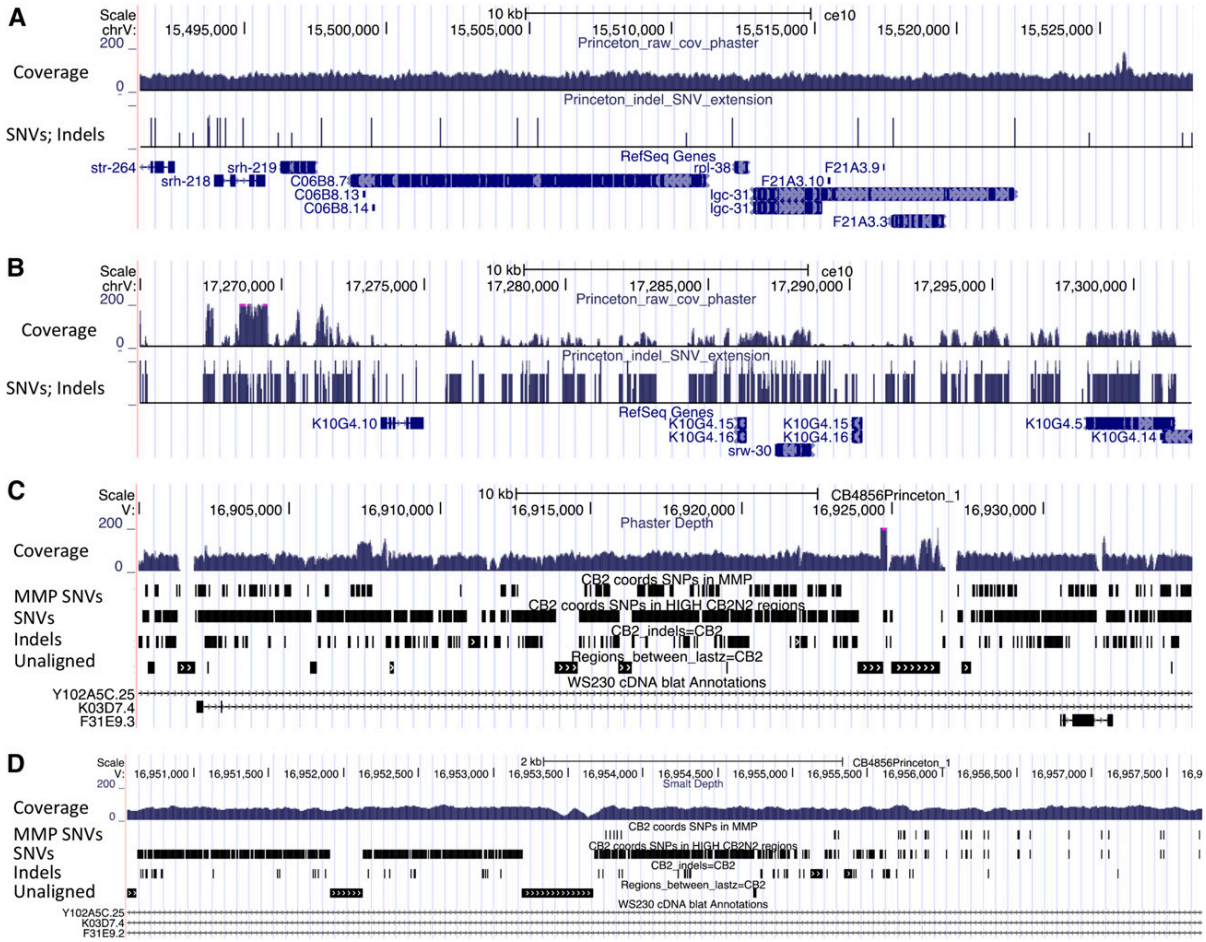

Figure 2 Read coverage and SNV density in the N2 reference genome and the iteratively corrected CB4856 genome. (A) A typical region for most of the genome is shown, with good coverage (top track) and infrequent SNVs and indels (second track). Genes are shown below. (B) A region of the N2 reference showing poor coverage and a high SNV/indel density with the Hawaiian reads. (C) After 20 iterations of reference-guided corrections, the same region as in B now has improved coverage by the CB4856 reads. In addition to coverage, the tracks show the SNV calls (MMP SNVs) reported in (Thompson *et al.* 2013), the SNV calls based on the new reference (SNVs), indels based on the new reference (Indels), and regions that failed to align with sequence present in the N2 reference (Unaligned). Gene models for each region are shown below. (D) The boundary of a divergent region (left) with a less divergent region of the genome is shown. The density of SNVs and indels changes abruptly. Tracks are as in C.

the corresponding Hawaiian chromosome. At least partial alignments were obtained for 26,571 of the 26,769 genes.

Validation of CB4856 by random long (Sanger) reads—SNVs

The CB4856 Sanger sequencing reads from the Washington University (random genomic short inserts) and the University of British Columbia (insert end sequences of a genomic fosmid library) projects were aligned against the final CB4856 reference using bwa (Li and Durbin 2009; Li *et al.* 2009). To assess whether the Hawaiian SNVs presented here were confirmed by those long reads, all reads with a mapping quality of >40 (25,667 total reads with 13,835,952 bases having phred quality scores ≥40) were retained for analysis. A total of 28,203 of the Hawaiian SNVs presented here were covered with at least one of the base calls from the long read set of phred quality ≥40. Of those 28,203, there were 27,680 (98.1%) where the long read confirmed the Hawaiian SNV. Of those, 8380 were in a divergent region, and 8123 of those were confirmed (96.9%). For the 19,823 not in divergent regions, 19,556 were confirmed (98.7%). For the 523 positions where the Hawaiian SNV was not confirmed by the long read data, we examined the basecall from the JR-Assembler for that position in the genome. Of the 523 positions, there were 416 total (172 in the divergent regions) where the position was in a JR-Assembler contig that was not one of the 221 complete JR-assembled contigs inserted into the final reference sequence. Small parts of other contigs were also incorporated into the final assembly, but because they were small, they were not excluded here. In 415 of the 416 cases, or less than half a percent of the cases, the base called in the JR-Assembler contig agreed with the Hawaiian SNV call. The single case that did not agree was in one of the divergent regions. The consistency between the

integrated assembly and the JR-contigs suggests that many of the differences with the fosmid end sequences could be due to strain differences or from errors in the Sanger reads.

Validation of CB4856 by random long (Sanger) reads—indels

There are 190 reciprocal indels where a bwa alignment suggests a long Sanger sequencing read alignment spanning a position of a reciprocal deletion (106 of which are in divergent regions). In each case, the reads involved were realigned with the CB4856 genome using cross_match (P. Green, unpublished data), providing an alternative alignment method and a full Smith–Waterman alignment of the read against the genome, and analyzed. Of those 190, in 3 (1.6%) cases (1 of the 106 in the highly divergent regions) the alignment of the long read suggests an alternate indel from what is in the CB4856 assembly.

There are 4900 simple insertions (1196 in highly divergent regions) where a bwa alignment identified a read that aligns within the simple insertion. After alignment with cross_match, in 11 cases (1%) the long read suggests an alternate insertion [4 (0.3%) in highly divergent regions].

Results

Generating a Hawaiian reference sequence

To develop a reference sequence for the CB4856 genome that would account for regions of poor representation and incorporate sequence present in CB4856 and missing in N2, we exploited data sets from across the world (Table 1). The missing and poorly defined regions of the CB4856 genome might reflect problems with specific libraries or sequencing platforms. Aligning sequence reads against the N2 reference genome from three independently generated libraries sequenced with the Illumina platform gave broadly similar patterns of coverage, with similar regions of poor or no coverage of the N2 reference. Calling SNVs on these data sets yielded essentially the same set of SNVs with no more than 3% of SNV calls unique to any one library (Figure S5). These regions of poor coverage persisted with different aligners that handle repeats differently—*phaster*, which places all reads that match equally well at multiple copies of identical repeats at the first copy (P. Green, personal communication), and bwa or *smalt* (http://sanger.ac.uk/resources/software/smalt/), which distributes the matching reads randomly among the copies. The small percentage of SNVs that were unique to individual data sets tended to be in regions where read coverage was marginal (*e.g.*, Figure 2), leading to above-threshold calls in one data set, but not in the others. Similar regions of poor coverage were also seen with the WGS library sequenced with SOLiD technology. These regions were also flagged with tools designed to detect alignment irregularities, such as REAPR and ALE (Clark *et al.* 2013; Hunt *et al.* 2013). Because these regions of poor coverage persisted across libraries and platforms, we postulated that these regions might reflect segments of the genome with much higher-than-average sequence differences that led to variability in read alignment or even alignment failure.

Table 2 Comparison of reference sequence lengths

Chromosome	N2	HA	Difference	%
I	15,072,423	14,890,789	181,634	1.21
II	15,279,345	14,885,952	393,393	2.57
III	13,783,700	13,596,826	186,874	1.36
IV	17,493,793	17,183,857	309,936	1.77
V	20,924,149	20,182,852	741,297	3.54
X	17,718,866	17,537,347	181,519	1.02
Total	100,272,276	98,277,623	1,994,653	1.99

% difference in size expressed as a percentage of the length of the chromosome in the Hawaiian genome.

To improve read alignment in such regions and to extend coverage, we initially employed a technique similar to the reference-guided assembly strategy used in *Arabidopsis* (Gan *et al.* 2011; Schneeberger *et al.* 2011). This guided assembly approach allowed us to exploit the continuity and high quality of the N2 reference sequence and to avoid the pitfalls associated with whole-genome assemblies using current technologies and algorithms. We used the called SNVs to change the N2 reference sequence to reflect the presumptive Hawaiian sequence throughout the genome (Figure 1). We also deleted N2 sequence where read coverage was largely lacking and split reads clearly spanned from one N2 region to another. We added sequence at the edge of insertion sites, based on the consensus sequences of the unmatched portions of reads. We then realigned the reads against the revised sequence. For these purposes, we used only a single data set to avoid possible differences between starting strains and to avoid the excessive computational demands from using all the WGS data sets. We selected the Princeton data set because it had very deep coverage (~70×), longer paired reads with larger inserts, and a high fraction of reads aligning to the genome. An initial cycle significantly improved alignments, and after 20 cycles of alignment and correction most regions had excellent agreement between the draft reference and the sequence reads. For example, alignment of CB4856 reads against the draft reference produced only 13 SNVs and 8 indels (compared with 219,787 SNVs and 46,674 indels against the N2 reference). As a broader measure of the improvement, average ALE scores per base, adjusted using N2 reads against the N2 reference as a baseline, decreased from −6.10 to −2.49 (Figure S6).

Even after 20 cycles of replacements, we found ~50–70 regions of 10–100 kb where the initial uneven coverage of the N2 sequence had nucleated improved coverage after the iterative correction/extension approach, but overall coverage of the region remained discontinuous; ALE and REAPR scores remained high in these regions (Figure 2). These regions often had indications of multiple deletions that lacked spanning reads to define their end points. As a result, they had not been removed in our iterative alignment and

Table 3 Number of deletion events and base counts in deletions in N2 and CB4856

	Deletions in N2		Deletions in CB4856	
Chromosome	Events	Bases	Events	Bases
I	5,693	94,543	6,158	233,930
II	7,370	230,813	7,692	478,884
III	5,464	99,324	6,008	275,997
IV	5,700	116,249	5,841	265,504
V	10,902	343,640	11,740	854,733
X	3,453	37,442	3,507	156,819
Total	38,582	922,011	40,946	226,5867

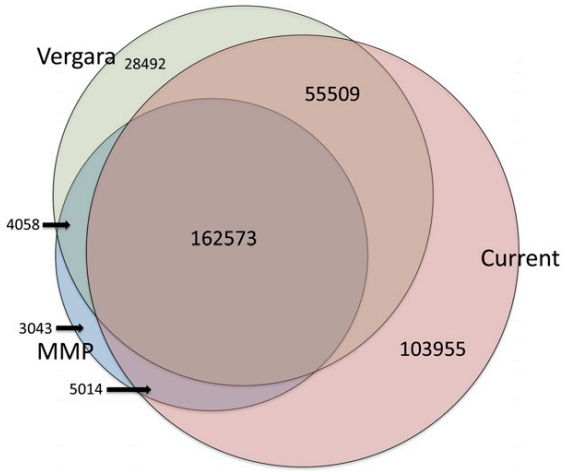

Figure 3 Overlap with previous SNV calls. A Venn diagram shows the overlap of the previous SNV calls with those obtained with the CB4856 reference.

correction procedure. To resolve these regions, we exploited a new whole-genome assembly algorithm, the JR-Assembler (Chu *et al.* 2013), which compares favorably to current assemblers in terms of quality, efficiency, and memory. We used the JR-Assembler to create a *de novo* assembly of the CB4856 genome. To reduce the number of likely false joins, we analyzed the assembled JR contigs with REAPR, splitting them where REAPR signaled problems.

To utilize the JR contigs, we first aligned the resultant 14,167 JR contigs, totaling 93,075,504 bases with an N50 of 15,785 bp (1601 contigs) against both the N2 and the iteratively improved genomes. The JR contigs aligned against N2 showed base-pair disagreements consistent with the called SNVs in most regions of the genome, *i.e.*, those that had good coverage with the CB4856 sequence reads. However, they revealed additional SNVs and indels in the 50–70 regions of spotty coverage (Figure 2). In contrast, when the JR contigs were aligned against the recursively corrected genome, they showed only rare differences, except in the regions with spotty coverage. In those regions, the JR contigs indicated additional base-pair changes and likely indels, suggesting that substituting the JR-contig sequences in the divergent regions of the recursive genome could be used (1) to remove some deletions that had failed to meet our criteria for removal in the iterative process, (2) to add sequence that had not been fully added in the 20 cycles, and (3) to correct some remaining substitutions (Figure 1 and Figure 2).

To confirm that the JR contigs were properly placed within the genome, we examined coverage metrics for each contig across a collection of 60 introgression lines and 49 RILs. For contigs aligning to these divergent regions, normal coverage with Hawaiian-derived reads was limited to those strains with a Hawaiian segment for that region, thus locating the contig within a region of a few megabases (Figure S7). Similar metrics were used to confirm placement of large insertions. Using only confirmed contigs, we replaced sequences in the recursive genome with the sequences from the JR contigs totaling >2 Mb (see *Materials and Methods* for details). We also removed presumptive deleted regions that had no JR contigs aligning and no aligning reads. These changes dropped the adjusted genome-wide ALE scores from −2.48 to −1.18 (Figure S6), and visual inspection of the divergent regions also supported a much-improved version of the genome (illustrated in Figure 1). Alignment of 42,261 Sanger reads from fosmid ends and plasmids (Table 1) also confirmed the substitutions and indels introduced in new CB4856 reference (Supporting Information).

From this combination of approaches we estimate that the resulting Hawaiian (HA) genome sequence totals 98,277,623 bases (98,291,416 including mitochondrial DNA), almost 2 Mb less than the N2 genome. All the CB4856 chromosomes are smaller than their N2 counterparts with II and V showing the greatest reductions at 2.57 and 3.54%, respectively (Table 2). The CB4856 genome size may be an underestimate because we have not been able to separate distinct copies of some repeats and some novel CB4856 sequences could still be missing. To provide a rough estimate of the total bases that might be missing from the CB4856 genome, we compared all the JR contigs to the CB4856 reference. There are only 22 assembled contigs (16,206 bases) that fail to align at all with our parameters. However, portions of other contigs fail to align to the CB4856 reference, and these total 1.12 Mb (2617 contigs). In addition, portions of other contigs (810 contigs, 1.74 Mb total) match the reference poorly (>1% of bases mismatched/inserted/deleted). Inspection of a sample of these contigs suggests that they could represent small insertions/deletions within the aligned sequence, duplicated sequence, poorly assembled sequence, or segments in the CB4856 with remaining problems. REAPR and ALE analysis also show some remaining problematic areas in the CB4856 reference, including high read coverage in regions that suggests unresolved duplications (sometimes overlapping more than one JR contig with poor alignment scores), low coverage, and inconsistent read pairing. Nonetheless, the CB4856 reference sequence

Table 4 SNV distribution

Chromosome	Total	Centers	SNV/kb centers	Arms	SNV/kb arms	% on arms
I	36,192	9,424	1.15	26,768	3.89	73.96
II	65,592	8,843	1.11	56,749	7.80	86.52
III	38,938	5,429	0.78	33,509	4.94	86.06
IV	36,198	9,016	1.00	27,182	3.20	75.09
V	129,096	18,179	2.02	110,917	9.30	85.92
X	21,034	8,076	1.08	12,958	1.27	61.61
Total	327,050	58,967	1.21	268,083	5.20	81.97

provides a much more complete picture of the genome for the analysis of variation.

Summary of sequence variation

To assess the sequence differences between the Hawaiian and N2 reference genomes, we aligned the CB4856 genome to N2 using LASTZ, a program that is able to perform pairwise, gapped, full chromosome-to-chromosome alignments even with quite distantly related sequences (Harris 2007). This alignment yielded a total of 327,050 SNVs and 79,529 indel events (File S1, File S2, File S3, File S4). In addition, 816 segments with sequence in both genomes failed to align with LASTZ and also failed to produce consistent alignments with different Smith–Waterman alignment implementations. Because these sequences fail to align, we cannot call variants in them. Inspection of several of these regions suggests several different origins. Some are extremely divergent and could represent overlapping deletions of an ancestral sequence, leaving unrelated sequences in the same place in the two present-day genomes. Others may just be of slightly greater divergence than LASTZ tolerates. Still others involve short sequences in one genome opposite a much larger region in the other genome that might be associated with repair events around a deletion. Because we were uncertain of the origin of the individual events, for purposes of analysis we treated the 816 segments as reciprocal insertion/deletion events.

If we treat the indels as deletions in N2 or CB4856 from a larger ancestral genome, we find similar numbers of deletion events in each genome, with small deletions (≤5 bases) accounting for 28,779 and 28,776 events in each N2 and CB4856, respectively. But the larger deletions (>5 bases) are slightly more numerous in CB4856 (12,181 *vs.* 9803), and the overall the number of bases deleted in CB4856 is much larger (2.2 Mb of ancestral genome lost in CB4856 *vs.* 0.92 Mb in N2). Chromosomes II and V have larger numbers of bases deleted. In addition to the indels, the 816 reciprocal insertion/deletion events show a similar trend of greater loss of sequence in the CB4856 genome, with these segments containing 458,526 bases in CB4856 and 1,109,331 bases in N2. *In toto* from the indels and reciprocal indels, N2 has 3.3 Mb of sequence not represented in CB4856, and CB4856 has 1.4 Mb of sequence not present in N2 (Table 3).

Among the deletions and insertions were copies of the transposons Tc1, Tc2, Tc3, Tc4, and Tc5. For example, among the 32 copies of Tc1 in the N2 reference, only 19 were detected in the CB4856 genome (Table S4). However, we found evidence for Tc1 copies in the CB4856 not present in N2. Because of the repeated nature of the sequence, the copies were incomplete in the CB4856 genome, but in 12 cases, Tc1 end sequences were present on both ends flanking a gap and, in another five cases, new Tc1 sequences were detected at one end. Differences in other Tc family transposons were also noted.

Our 327,050 SNV calls detect most of the SNVs reported earlier (Thompson *et al.* 2013; Vergara *et al.* 2014), but we find almost twice as many SNVs as Thompson *et al.* (2013) and 103,955 not reported by Vergara *et al.* (2014) (Figure 3). These novel SNVs are supported by conventional Sanger sequenced reads at a rate similar to those found in common by all three reports (Supporting Information). Many of the SNVs reported in Vergara *et al.* (2014) but not found here are adjacent to small deletions associated with homopolymer runs not reported in Vergara *et al.* (2014) [these authors report only 31,791 indels, compared with 46,544 detected in the Million Mutation Project (Thompson *et al.* 2013) and almost 80,000 here], suggesting that some of their called SNVs result from alignment issues and/or problems in runs (Becker *et al.* 2012).

Next, we looked at the distribution of the SNVs and indels across and within the chromosomes. *C. elegans* chromosomes have a distinctive organization, with the outer 20–30% of each chromosome (the arms) exhibiting a higher rate of recombination and a higher fraction of repeated sequences (Barnes *et al.* 1995; Rockman and Kruglyak 2009). They also contain the bulk of genes for large, rapidly evolving gene families. Consistent with previous reports, we find the number of SNVs is higher on the autosome arms than in the centers (Table 4, File S4, Figure S8), with as much as 86% of SNVs on arms . The distribution of indel events follows a similar pattern.

More strikingly, we noted a strong clustering of variants in smaller regions where we had extended sequence from the ends of aligned segments during the iterative alignment process (Figure 1A) and where we had replaced sequence with the JR-assembled contigs (Figure 1B). To detect these regions of higher divergence systematically, we clustered

Table 5 SNVs in divergent regions

	Divergent regions			Other regions		
Chromosome	SNVs	Bases	SNVs/kb	SNVs	Bases	SNVs/kb
I	3,940	87,170	45.20	32,252	14,985,253	2.15
II	27,649	709,991	38.94	37,943	14,569,354	2.60
III	13,962	344,847	40.49	24,976	13,438,853	1.86
IV	5,657	206,442	27.40	30,541	17,287,351	1.77
V	77,704	1,444,451	53.79	51,392	19,479,698	2.64
X	900	38261	23.52	20,134	17,680,605	1.14
Total	129,812	2,831,162	45.85	197,238	97,441,114	2.02

1-kb windows with >1.4% bases different or >500 bases deleted or inserted, retaining clusters of ≥9 kb. We manually reviewed each of the resulting clusters, adjusting the end points to more precisely reflect increasing SNV density, merging closely spaced clusters separated by repeats or other regions where SNVs were unable to be called. This procedure produced 61 regions containing a strikingly high proportion of SNVs and spanning 2,313,859 bases in the CB4856 sequence and 2,831,162 in the N2 genome (Table 5; see also Table S3 for a list of regions with coordinates and Table S5 for SNV counts in the absence of the divergent regions). The boundaries between these divergent regions and other regions are usually very sharp (Figure 2 and Figure 4). The segments are scattered across all six chromosomes, range in size from 8 to >162 kb and show 2–15.8% sequence divergence from N2 without an obvious correlation between size and divergence (Figure 5). The autosomal clusters fall principally on the arms with just one cluster in the central region of IV and three clusters in the central region of V (Figure S8). The X chromosome has only two clusters, and these have just 2 and 4.6% divergence each. The autosomal divergent regions include the *peel-1 zeel-1* region and the *glc-1* gene, both of which had been previously reported as having an elevated sequence divergence (Seidel *et al.* 2011; Ghosh *et al.* 2012).

Curiously, in comparing our results with prior studies using array CGH (Maydan *et al.* 2007; Maydan *et al.* 2010), we find that more than one-third of their deletion calls fall within these divergent regions, often extending across most of the region. Apparently the sequence divergence within the regions led to poor hybridization with the probes and resultant scoring of the area as deleted. The remaining arrayCGH deletions overlap extensively with deletions in the CB4856 reference except in one case [WBVar00091092; niDf71(III); chrIII:13778179–13781358] where we have normal coverage throughout the region.

Functional impact of the sequence variants

Prediction of the functional effects of the variants using WS230 annotation shows that a large portion of the protein-coding potential of the Hawaiian genome is altered (Table 6; File S5). Across the whole genome, 8140 (40%) of the 20,504 protein-coding genes have some coding change (for complete lists of the alterations, see File S2, File S3, and File S5, File S6). Of these genes, 1885 protein-coding genes have a likely loss-of-function (LOF) mutation—an induced stop codon, a frameshift, or a deletion across a splice junction—and 357 of these delete the gene entirely. Our reference also detects previously reported variants throughout the genome (Table S6).

In the 2.83 Mb of divergent regions, the relative impact of the variation on protein-coding genes is even greater. Of the 883 genes in these regions, 866 (98%) have some coding change, with 576 genes having LOF changes and 195 genes deleted entirely. Those genes with LOF changes could all be pseudogenes in CB4856, perhaps as a rapid means of adaptation (Olson 1999). However, the persistence of these regions over long evolutionary times (see below) suggests that they contain functionally important genes.

The genes altered by the LOF variants are disproportionately composed of members of large, rapidly evolving gene families, including the *math*, *bath* (btb-math), *clec* (c-lectin), *fbx* (F-box), and seven-transmembrane serpentine receptor genes. By contrast, the *nhr* (nuclear hormone receptor) genes are not overrepresented in LOF variants. The same large, rapidly evolving gene families are also heavily represented in the divergent regions. The *math* family is notable, having 37 of its 49 members in the divergent regions. Within the divergent regions the *math*, *bath*, and *fbx* families also suffer a disproportionate number of LOF variants. The members of these families in the divergent regions that do not contain LOF variants also show a d_N/d_S ratio (ratio of divergence at nonsynonymous and synonymous sites) approaching or even exceeding 1. By contrast, the seven-transmembrane and *nhr* gene families are relatively spared. Members of these families without LOF variants also show a lower fraction of nonsynonymous changes than the *math* and *fbx* families.

Some genes with easily visible mutant phenotypes contain LOF variants, such as *ced-1*, *ced-6*, *unc-13*, and *unc-49*. However, inspection of the variants in these genes in comparison with WormBase models and the extensive modENCODE RNA-seq data (Gerstein *et al.* 2014) suggest in each case that the annotation likely requires correction. For example, the putative stop codon in the *unc-13* model falls in a splice form that has no RNA-seq support; instead, the RNA-seq evidence suggests that the exon in question is an alternate first exon and the "nonsense" change lies in the 5′ UTR. Similarly, the putative frameshift mutation (an insertion of

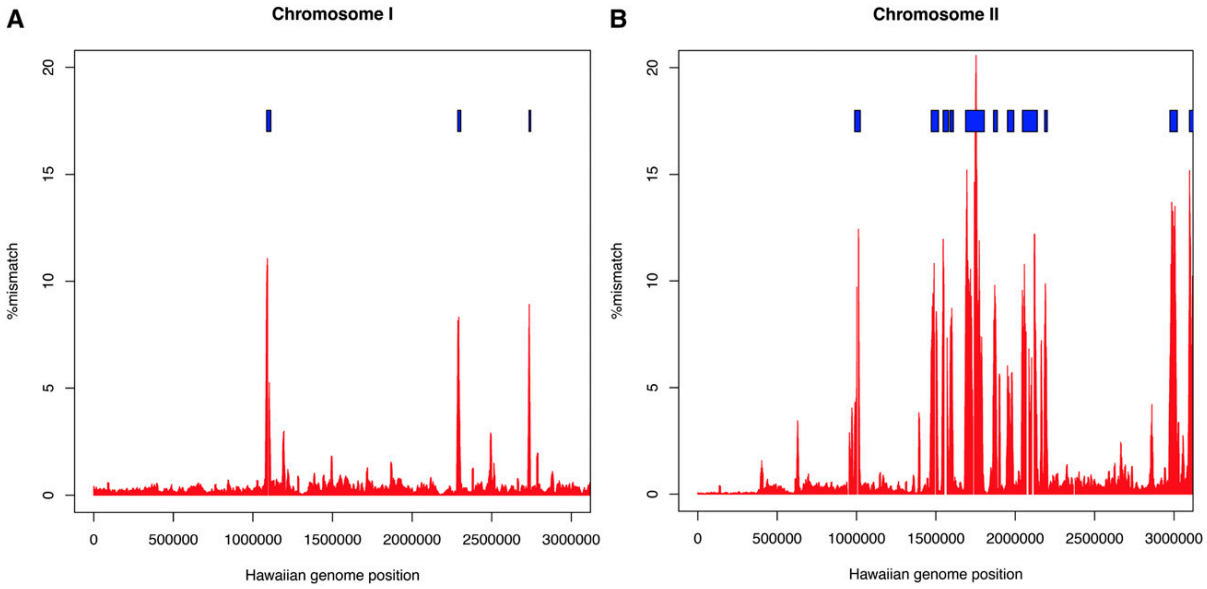

Figure 4 Density of variant sites in the first three megabases of (A) chromosome I and (B) chromosome II. Blue boxes indicate the regions identified as highly divergent.

a G) in *ced-6* falls adjacent to a noncanonical splice donor sequence in the WormBase gene model, where the presence of an extra G would allow the use of a canonical GT donor. Inspection of data from an N2 strain (VC2010) suggests that a G is missing here in the N2 reference.

Other regions that include genes such as *tbx-30*, *tbx-31*, *vit-3*, and *vit-4* have suffered large deletions in the CB4856 genome, with *tbx-30* deleted entirely. However, in the N2 genome *tbx-30* has an exact duplicate, *tbx-42*, as part of an inverted duplicated segment; presumably, the loss of a single copy is tolerated. The two *vit* genes are adjacent in the genome, and in this case the deletion appears to fuse the two genes into a single gene.

Although genes have been lost in the CB4856 genome, genes may also be present in CB4856 that are defective or absent in N2. To look for these genes, we generated gene models by using an *ab initio* gene prediction tool, Genefinder, and by aligning CB4856 RNA-seq data (Table 1) against the CB4856 reference (Trapnell *et al.* 2012). Among the recent N2 pseudogenes that arose by duplication, we found two gene models in CB4856 that had full opening reading frames that were similar to the parent genes. In sequences inserted in CB4856 relative to N2, Genefinder produced models that were often supported by RNA-seq data and had similarity to members of multigene families. For example, two large insertions of 5 and 15 kb in one divergent region (II:1544781–1579478 CB4856 coordinates; II:1613411–1636156 N2 coordinates) contain seven Genefinder models, all of which have similarity to *fbx* genes. Similarly, in regions where LASTZ failed to align the N2 and CB4856 sequence, there were gene models with RNA-seq support in the CB4856 sequences. For example, in two adjacent regions that fail to align by LASTZ in a divergent region (II:2974336–3020963 CB4856 coordinates; II:3197502–3250769 N2 coordinates), there are two gene models, one with similarity to seven-transmembrane receptors and the second to *fbx* genes. Thus, while CB4856 has lost genes present in N2, it has also gained genes.

We also examined genes present in both species with multiple differences that suggested that they were inactive in CB4856. One gene—*F47H4.2*—stood out because of the multiple changes reported from the LASTZ alignments, including 691 SNVs across the two isoforms that, considered together, result in 511-amino-acid substitutions, 1 nonsense mutation, and 13 frameshifting indels. Despite these multiple variants, a Genefinder model in the region of CB4856 predicts a protein of 628 amino acids with its seven exons having open reading frames similar in length to those in N2 and with six of those having splice junctions in precisely equivalent places (Figure S9). The predicted nonsense codon is present in an unused frame, a consequence of flanking, compensating frameshift variants. The 628-amino-acid predicted protein, when evaluated by blastp (Altschul *et al.* 1990) against the N2 proteins, yielded a match that covered amino acids 1–633 (P = 2.7*e*-127) of an *F47H4.2* isoform, spanning two FTH domains that are also found in F-box genes. The Genefinder exons also are matched by RNA-seq reads. Open reading frames also exist in CB4856 for the final three exons of *F47H4.2* and are partially incorporated into a second Genefinder model (Figure S9). These results suggest that, rather than containing an inactivated gene, the CB4856 region encodes a homologous protein. Like *F47H4.2*, other genes that appear to have suffered a LOF variant in CB4856 also have Genefinder models in syntenic regions

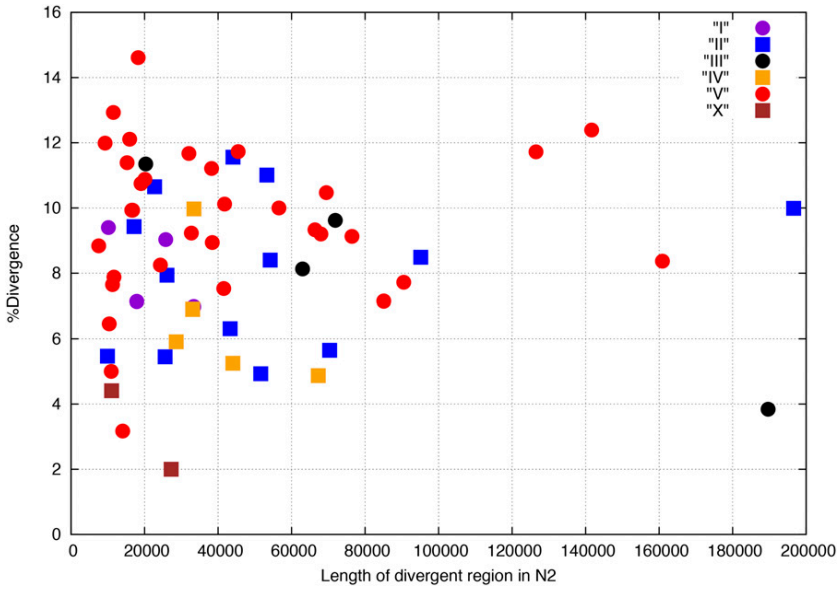

Figure 5 Percent divergence by length of divergent region per chromosome. The mutational events (SNVs and indels, counting each indel as a single event) per aligned bases (percentage divergence) are plotted for each region against the length of the region in N2. The chromosomal assignment for each region is indicated in the inset.

that are often supported by RNA-seq data and may similarly encode full-length proteins.

Divergent regions in other C. elegans strains

Prior studies of the *glc-1* and *peel-1 zeel-1* regions showed that both CB4856 and N2 haplotypes were widely distributed in strains recovered from the wild (Seidel *et al.* 2011; Ghosh *et al.* 2012). To determine haplotype representation of the 61 divergent regions in other wild strains, we exploited the MPS sequence reads acquired previously (Thompson *et al.* 2013) for each of 39 strains and aligned them against the N2 reference and the CB4856 reference. We expected that, if a particular region in a given strain had the N2 haplotype, its reads would align well with the N2 reference but poorly with the CB4856 reference and vice versa. To assess the alignment quality, we calculated ALE scores using the N2 and CB4856 reads aligned against each other as controls.

Using the ALE scores (Figure S10), we cataloged the regions as N2-like, CB4856-like, intermediate, or different from either. The results (Figure 6) for the 44 regions giving consistent scoring show that, while, overall, N2-like haplotypes are most frequent, the CB4856-like sequence is found in at least one other strain for all but six of the regions. For five regions, the CB4856-like haplotype is predominant. Two regions are represented mainly by sequence that matches neither N2 nor CB4856 well, suggesting the presence of another version of the sequence in these strains. Other regions match both N2 and CB4856 at intermediate levels, *e.g.*, V:18193641–18260001, and inspection of these regions suggests that some have one segment that matches CB4856 well and a second segment that matches N2 well, perhaps reflecting recombination events between the two haplotypes. Most of the strains have unique combinations of sequences; however, JU1171, MY2, and MY14 all share the same pattern (both MY2 and MY14 were isolated from Munster, Germany; share 96% of SNV calls and share 96% with JU1171, isolated in Chile; they likely represent a single isotype), and ED3057 and ED3072 are similar (the latter two were both isolated in Kenya and share 97% of their SNV calls and also likely represent a single isotype).

Discussion

We have used a variety of resources and methods to produce a draft reference CB4856 genome sequence. We combined iterative alignment of deep whole-genome sequence reads to a progressively corrected CB4856 reference version with a *de novo* whole-genome assembly. We used the assembly assessment tools ALE and REAPR to monitor progress, to identify problem areas needing improvement, and to break suspicious joins in the *de novo* assembly. Sequence reads from ILs and RILs helped guide placement of the *de novo* contigs into the reference. Sanger reads from fosmid-insert ends and random clones were critical in validating the reference sequence at each step.

The resulting CB4856 reference sequence extends and refines the scope of the variation between N2 and CB4856 from prior studies. In particular, the draft sequence reveals 61 regions of substantially higher divergence than the rest of the genome. These regions total 2.8% of the N2 genome and contain 40% (129,812/327,050) of total SNVs and 21% (16,822/79,558) of total indels. A survey of genome sequence data from other wild isolates of *C. elegans* shows that generally in these regions they closely resemble one or the other genome. However, some isolates appear to have regions that are divergent from both N2 and CB4856 and in segments outside the divergent regions some strains show divergence from both N2 and CB4856. These findings

Table 6 Genes in diverged regions and with LOF mutations

	Genome			Divergent regions			
Gene class	Total	Disabling	Expected[a]	Total	Expected	Disabling	Expected[b]
Serpentine receptor (*sr**/*str*)	1346	204	123.7 2.20e-13[c]	118	49.7 7.38e-14	72	80.0 8.70e-1
F-box (*fbx**)	353	129	32.5 1.56e-45	71	15.3 3.75e-28	60	46.3 1.51e-4
C-lectin (*clec*)	254	53	23.4 1.05e-8	31	11.0 1.81e-7	19	20.2 7.49e-1
Math (*math*)	48	39	4.4 1.92e-32	37	2.1 2.00e-41	34	24.1 1.47e-4
Bath (*bath*)	37	16	3.4 4.83e-8	15	1.6 1.11e-11	14	9.7 1.42e-2
Nuclear hormone receptor (*nhr*)	278	26	25.6 4.90e-1	27	11.8 7.38e-05	10	17.6 9.99e-1

[a] The expected number of disabled genes in the total genome based on 20,504 genes and 1885 disabled overall.
[b] The expected number of genes in the divergent regions based on 883 genes of the 20,504 genes in the genome and 576 of the 883 genes disabled.
[c] Hypergeometric test.

suggest that N2 and CB4856 do not capture all the divergent haplotypes in these regions or even all the regions with divergent haplotypes extant in this species. The reference genome, associated data sets, and annotations can be viewed through a track data hub listed on the University of California at Santa Cruz browser (http://genome.ucsc.edu/cgi-bin/hgHubConnect and connect to "*C. elegans* isolates").

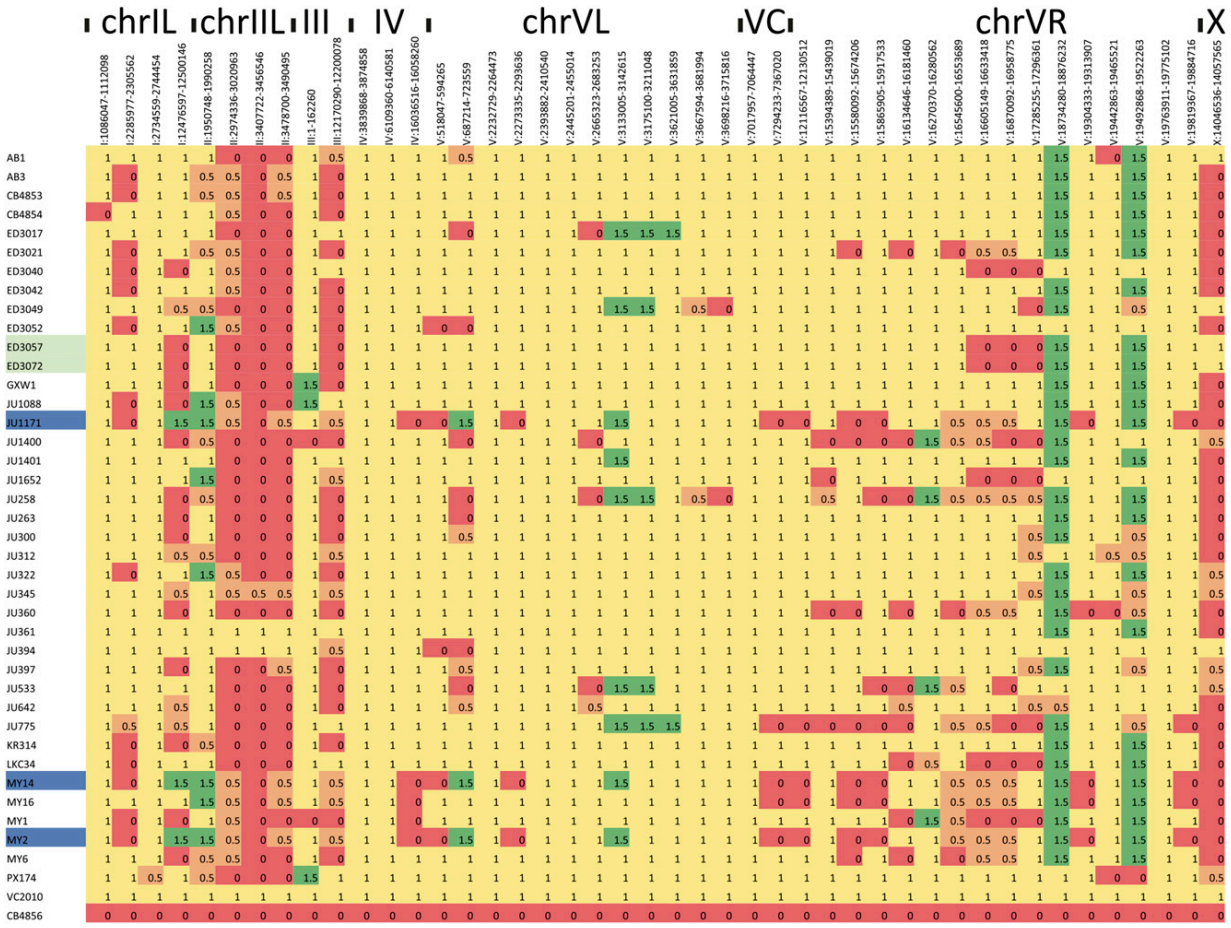

Figure 6 A heatmap representation of the allelic content of the 39 strains (rows) across 44 of the 61 divergent regions (columns). Regions matching N2 (yellow) and CB4856 (red) are indicated along with intermediate regions (orange) and regions different from either (green). For reference, an N2-derived strain, VC2010, and CB4856 are shown in the bottom two rows. Strains that may represent the same isotype are highlighted in blue and green.

The high level of divergence between the two haplotypes of N2 and CB4856 in these regions indicates an ancient origin. Possible explanations could include long-term balancing selection, as suggested for the *glc-1* haplotypes (Ghosh *et al.* 2012), that were estimated to have arisen $\sim 7 \times 10^6$ generations ago. Applying the same methods to the 61 regions described here gives an overall estimate of d_S (average number of nucleotide differences between sequences per synonymous site) of 0.12 and similar estimate of 6.7×10^6 generations. The male–female species *C. remanei* shows 4% divergence in wild populations (Dey *et al.* 2012). Perhaps these divergent regions are a remnant of the variation present at the conversion to a hermaphroditic species, with all 61 regions maintained by balancing selection since the origin of the species. Alternatively, the sequences may have evolved independently in different, isolated populations, followed by one or more subsequent cross-breeding events, leading to admixture and introgression. Regardless of the origin, haplotypes for only a few regions have persisted, perhaps maintained by balancing selection. Although the high degree of sequence divergence might be expected to interfere with crossing over in the initial hybrid, resulting in a severe lack of fitness, nematodes with this level of divergence have been found to be fertile (Dey *et al.* 2013). Perhaps the need for only one recombination event per chromosome makes the *Caenorhabditis* species more tolerant of long tracts of divergent sequences. Regardless of the nature of the original event(s), balancing selection is likely to be playing a role in the maintenance of these regions across the globe.

What is driving the balancing selection? Genes that belong to large, rapidly evolving gene families and that have a putative role in the interaction with the environment are abundant in these regions (Thomas 2006; Thomas and Robertson 2008). The alternative haplotypes may harbor combinations of genes and alleles that provide selective advantage in the face of changing environmental conditions. The *glc-1* polymorphisms provide a specific example of polymorphisms that would confer selective advantage in an environment containing avermectin or related compounds. In the absence of avermectin, the defective *glc-1* may lead to a fitness disadvantage (Ghosh *et al.* 2012). But also, *glc-1* lies in a divergent region with multiple seven-transmembrane receptor and *clec* genes. Perhaps these other genes also have a role in balancing selection.

Beyond the revelation of these divergent regions, the CB4856 reference sequence provides investigators with a comprehensive list of the changes between CB4856 and N2 (File S1, File S2, File S3, File S5). These lists should prove useful to those investigators using wild isolates of *C. elegans*. For example, the failure of a probe to hybridize well to the CB4856-derived sequences might lead to false negatives. RNA-seq biases introduced by mapping to the N2 reference may also distort any signal. These biases are particularly relevant for the divergent regions. The extensive catalog of differences now available between these two strains in combination with the powerful genetic approaches available in *C. elegans* should facilitate the dissection of the growing number of phenotypic differences.

Our findings also have broader implications for studies comparing genome sequences to reference sequences using short reads. The high degree of sequence divergence in these divergent regions compromises alignment of short reads. Studies that use only standard alignment of reads to a reference fail to assess such divergent sequence. If similar areas of high divergence are present in other species, they too would be missed. Our results indicate the importance of going beyond simple read alignment in the assessment of variability between different genomes.

Acknowledgments

We thank Brent Ewing and Stephane Flibotte for many discussions on sequence alignment, assembly, and analysis; Yeh Teng Wen and the JR-Assembler team for help in installing and running their assembler; and Asher Cutter for ideas about the origins of the divergent regions. The work was supported by *The American Recovery and Reinvestment Act* Grand Opportunities (ARRA GO) grant HG005921 from the *National Human Genome Research Institute (NHGRI)*, by grant HG007355 from NHGRI, and by the William H. Gates Chair of Biomedical Sciences. L.B.S. was funded by The Netherlands Organisation for Scientific Research (project no. 823.01.001).

Literature Cited

Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman 1990 Basic local alignment search tool. J. Mol. Biol. 215: 403–410.

Andersen, E. C., J. P. Gerke, J. A. Shapiro, J. R. Crissman, R. Ghosh *et al.*, 2012 Chromosome-scale selective sweeps shape *Caenorhabditis elegans* genomic diversity. Nat. Genet. 44: 285–290.

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

Andersen, E. C., T. C. Shimko, J. R. Crissman, R. Ghosh, J. S. Bloom *et al.*, 2015 A powerful new quantitative genetics platform, combining *Caenorhabditis elegans* high-throughput fitness assays with a large collection of recombinant strains. G3 (Bethesda) 5: 911–920.

Barnes, T. M., Y. Kohara, A. Coulson, and S. Hekimi, 1995 Meiotic recombination, noncoding DNA and genomic organization in *Caenorhabditis elegans*. Genetics 141: 159–179.

Becker, E. A., C. M. Burns, E. J. Leon, S. Rajabojan, R. Friedman *et al.*, 2012 Experimental analysis of sources of error in evolutionary studies based on Roche/454 pyrosequencing of viral genomes. Genome Biol. Evol. 4: 457–465.

Cao, J., K. Schneeberger, S. Ossowski, T. Gunther, S. Bender *et al.*, 2011 Whole-genome sequencing of multiple *Arabidopsis thaliana* populations. Nat. Genet. 43: 956–963.

Capra, E. J., S. M. Skrovanek, and L. Kruglyak, 2008 Comparative developmental expression profiling of two *C. elegans* isolates. PLoS ONE 3: e4055.

C. elegans Sequencing Consortium, 1998 Genome sequence of the nematode *C. elegans*: a platform for investigating biology. Science 282: 2012–2018.
Chu, T. C., C. H. Lu, T. Liu, G. C. Lee, W. H. Li *et al.*, 2013 Assembler for de novo assembly of large genomes. Proc. Natl. Acad. Sci. USA 110: E3417–E3424.
Clark, S. C., R. Egan, P. I. Frazier, and Z. Wang, 2013 ALE: a generic assembly likelihood evaluation framework for assessing the accuracy of genome and metagenome assemblies. Bioinformatics 29: 435–443.
de Bono, M., and C. I. Bargmann, 1998 Natural variation in a neuropeptide Y receptor homolog modifies social behavior and food response in *C. elegans*. Cell 94: 679–689.
Dey, A., Y. Jeon, G. X. Wang, and A. D. Cutter, 2012 Global population genetic structure of *Caenorhabditis remanei* reveals incipient speciation. Genetics 191: 1257–1269.
Dey, A., C. K. Chan, C. G. Thomas, and A. D. Cutter, 2013 Molecular hyperdiversity defines populations of the nematode *Caenorhabditis brenneri*. Proc. Natl. Acad. Sci. USA 110: 11056–11060.
Doroszuk, A., L. B. Snoek, E. Fradin, J. Riksen, and J. Kammenga, 2009 A genome-wide library of CB4856/N2 introgression lines of *Caenorhabditis elegans*. Nucleic Acids Res. 37: e110.
Gan, X., O. Stegle, J. Behr, J. G. Steffen, P. Drewe *et al.*, 2011 Multiple reference genomes and transcriptomes for *Arabidopsis thaliana*. Nature 477: 419–423.
Ghosh, R., E. C. Andersen, J. A. Shapiro, J. P. Gerke, and L. Kruglyak, 2012 Natural variation in a chloride channel subunit confers avermectin resistance in *C. elegans*. Science 335: 574–578.
Gerstein, M. B., J. Rozowsky, K. K. Yan, D. Wang, C. Cheng *et al.*, 2014 Comparative analysis of the transcriptome across distant species. Nature 28: 445–448.
Harris, R. S., 2007 *Improved Pairwise Alignment of Genomic DNA*. Ph.D. Thesis, The Pennsylvania State University Press.
Hodgkin, J., and T. Doniach, 1997 Natural variation and copulatory plug formation in *Caenorhabditis elegans*. Genetics 146: 149–164.
Hunt, M., T. Kikuchi, M. Sanders, C. Newbold, M. Berriman *et al.*, 2013 REAPR: a universal tool for genome assembly evaluation. Genome Biol. 14: R47.
Kammenga, J. E., A. Doroszuk, J. A. Riksen, E. Hazendonk, L. Spiridon *et al.*, 2007 A *Caenorhabditis elegans* wild type defies the temperature-size rule owing to a single nucleotide polymorphism in tra-3. PLoS Genet. 3: e34.
Kent, W. J., R. Baertsch, A. Hinrichs, W. Miller, and D. Haussler, 2003 Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes. Proc. Natl. Acad. Sci. USA 100: 11484–11489.
Koch, R., H. G. van Luenen, M. van der Horst, K. L. Thijssen, and R. H. Plasterk, 2000 Single nucleotide polymorphisms in wild isolates of *Caenorhabditis elegans*. Genome Res. 10: 1690–1696.
Li, H., and R. Durbin, 2009 Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25: 1754–1760.
Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan *et al.*, 2009 The Sequence Alignment/Map format and SAMtools. Bioinformatics 25: 2078–2079.
Li, Y., O. A. Alvarez, E. W. Gutteling, M. Tijsterman, J. Fu *et al.*, 2006 Mapping determinants of gene expression plasticity by genetical genomics in *C. elegans*. PLoS Genet. 2: e222.
Mackay, T. F., S. Richards, E. A. Stone, A. Barbadilla, J. F. Ayroles *et al.*, 2012 The *Drosophila melanogaster* Genetic Reference Panel. Nature 482: 173–178.
Maydan, J. S., S. Flibotte, M. L. Edgley, J. Lau, R. R. Selzer *et al.*, 2007 Efficient high-resolution deletion discovery in *Caenorhabditis elegans* by array comparative genomic hybridization. Genome Res. 17: 337–347.
Maydan, J. S., A. Lorch, M. L. Edgley, S. Flibotte, and D. G. Moerman, 2010 Copy number variation in the genomes of twelve natural isolates of *Caenorhabditis elegans*. BMC Genomics 11: 62.
Nicholas, W. L., E. C. Dougherty, and E. L. Hansen, 1959 Axenic cultivation of *Caenorhabditis briggsae* (Nematoda, Rhabditidae) with chemically undefined supplements: comparative studies with related nematodes. Ann. N. Y. Acad. Sci. 77: 218–236.
Olson, M. V., 1999 When less is more: gene loss as an engine of evolutionary change. Am. J. Hum. Genet. 64: 18–23.
Perkins, J. D., 2010 *Comparison of Fosmid Libraries Made from Two Geographic Isolates of Caenorhabditis elegans*. M.Sc. Thesis, University of British Columbia, Vancouver.
Pollard, D. A., and M. V. Rockman, 2013 Resistance to germline RNA interference in a *Caenorhabditis elegans* wild isolate exhibits complexity and nonadditivity. G3 (Bethesda) 3: 941–947.
Rockman, M. V., and L. Kruglyak, 2008 Breeding designs for recombinant inbred advanced intercross lines. Genetics 179: 1069–1078.
Rockman, M. V., and L. Kruglyak, 2009 Recombinational landscape and population genomics of *Caenorhabditis elegans*. PLoS Genet. 5: e1000419.
Rockman, M. V., S. S. Skrovanek, and L. Kruglyak, 2010 Selection at linked sites shapes heritable phenotypic variation in *C. elegans*. Science 330: 372–376.
Schacherer, J., J. A. Shapiro, D. M. Ruderfer, and L. Kruglyak, 2009 Comprehensive polymorphism survey elucidates population structure of *Saccharomyces cerevisiae*. Nature 458: 342–345.
Schneeberger, K., S. Ossowski, F. Ott, J. D. Klein, X. Wang *et al.*, 2011 Reference-guided assembly of four diverse *Arabidopsis thaliana* genomes. Proc. Natl. Acad. Sci. USA 108: 10249–10254.
Schwartz, S., W. J. Kent, A. Smit, Z. Zhang, R. Baertsch *et al.*, 2003 Human-mouse alignments with BLASTZ. Genome Res. 13: 103–107.
Seidel, H. S., M. Ailion, J. Li, A. van Oudenaarden, M. V. Rockman *et al.*, 2011 A novel sperm-delivered toxin causes late-stage embryo lethality and transmission ratio distortion in *C. elegans*. PLoS Biol. 9: e1001115.
Snoek, L. B., K. J. Van der Velde, D. Arends, Y. Li, A. Beyer *et al.*, 2013 WormQTL: public archive and analysis web portal for natural variation data in *Caenorhabditis* spp. Nucleic Acids Res. 41: D738–D743.
Snoek, L. B., H. E. Orbidans, J. J. Stastna, A. Aartse, M. Rodriguez *et al.*, 2014 Widespread genomic incompatibilities in *Caenorhabditis elegans*. G3 (Bethesda) 4: 1813–1823.
Sterken, M. G., L. B. Snoek, J. E. Kammenga, and E. C. Andersen, 2015 The laboratory domestication of *Caenorhabditis elegans*. Trends Genet. 31: 224–231.
Stewart, M. K., N. L. Clark, G. Merrihew, E. M. Galloway, and J. H. Thomas, 2005 High genetic diversity in the chemoreceptor superfamily of *Caenorhabditis elegans*. Genetics 169: 1985–1996.
Stoeckius, M., D. Grun, and N. Rajewsky, 2014 Paternal RNA contributions in the *Caenorhabditis elegans* zygote. EMBO J. 33: 1740–1750.
Sulston, J. E., and S. Brenner, 1974 The DNA of *Caenorhabditis elegans*. Genetics 77: 95–104.
Swan, K. A., D. E. Curtis, K. B. McKusick, A. V. Voinov, F. A. Mapa *et al.*, 2002 High-throughput gene mapping in *Caenorhabditis elegans*. Genome Res. 12: 1100–1105.
Thomas, J. H., 2006 Adaptive evolution in two large families of ubiquitin-ligase adapters in nematodes and plants. Genome Res. 16: 1017–1030.
Thomas, J. H., and H. M. Robertson, 2008 The *Caenorhabditis* chemoreceptor gene families. BMC Biol. 6: 42.
Thompson, O., M. Edgley, P. Strasbourger, S. Flibotte, B. Ewing *et al.*, 2013 The million mutation project: a new approach to genetics in *Caenorhabditis elegans*. Genome Res. 23: 1749–1762.
Trapnell, C., A. Roberts, L. Goff, G. Pertea, D. Kim *et al.*, 2012 Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat. Protoc. 7: 562–578.

van der Velde, K. J., M. de Haan, K. Zych, D. Arends, L. B. Snoek *et al.*, 2014 WormQTLHD: a web database for linking human disease to natural variation data in *C. elegans*. Nucleic Acids Res. 42: D794–D801.

Vergara, I. A., M. Tarailo-Graovac, C. Frech, J. Wang, Z. Qin *et al.*, 2014 Genome-wide variations in a natural isolate of the nematode *Caenorhabditis elegans*. BMC Genomics 15: 255.

Vinuela, A., L. B. Snoek, J. A. Riksen, and J. E. Kammenga, 2012 Aging uncouples heritability and expression-QTL in *Caenorhabditis elegans*. G3 (Bethesda) 2: 597–605.

Volkers, R. J., L. B. Snoek, C. J. Hubar, R. Coopman, W. Chen *et al.*, 2013 Gene-environment and protein-degradation signatures characterize genomic and phenotypic diversity in wild *Caenorhabditis elegans* populations. BMC Biol. 11: 93.

Wicks, S. R., R. T. Yeh, W. R. Gish, R. H. Waterston, and R. H. Plasterk, 2001 Rapid gene mapping in *Caenorhabditis elegans* using a high density polymorphism map. Nat. Genet. 28: 160–164.

Communicating editor: M. Johnston