Whole-genome sequencing and variant discovery in *C. elegans*

LaDeana W Hillier, Gabor T Marth, Aaron R Quinlan, David Dooling, Ginger Fewell, Derek Barnett, Paul Fox, Jarret I Glasscock, Matthew Hickenbotham, Weichun Huang, Vincent J Magrini, Ryan J Richt, Sacha N Sander, Donald A Stewart, Michael Stromberg, Eric F Tsung, Todd Wylie, Tim Schedl, Richard K Wilson & Elaine R Mardis

Supplementary figures and text:

Supplementary Figure 1 Solexa single end read coverage levels mapped onto the *C. elegans* genome.

Supplementary Figure 2 Average N2 Bristol Solexa single end read coverage at different A+T percentages.

Supplementary Figure 3 Chromosomal distribution pattern of CB4858 SNPs and indels.

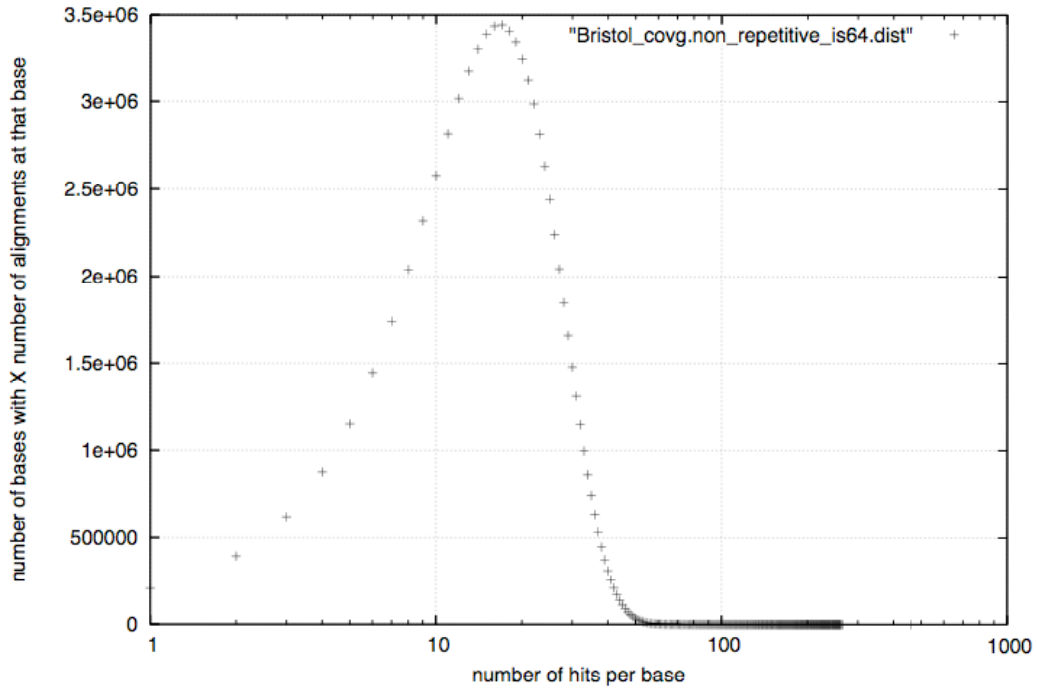
Supplementary Figure 4 Codon position bias in validated CB4858 SNPs.

Supplementary Data

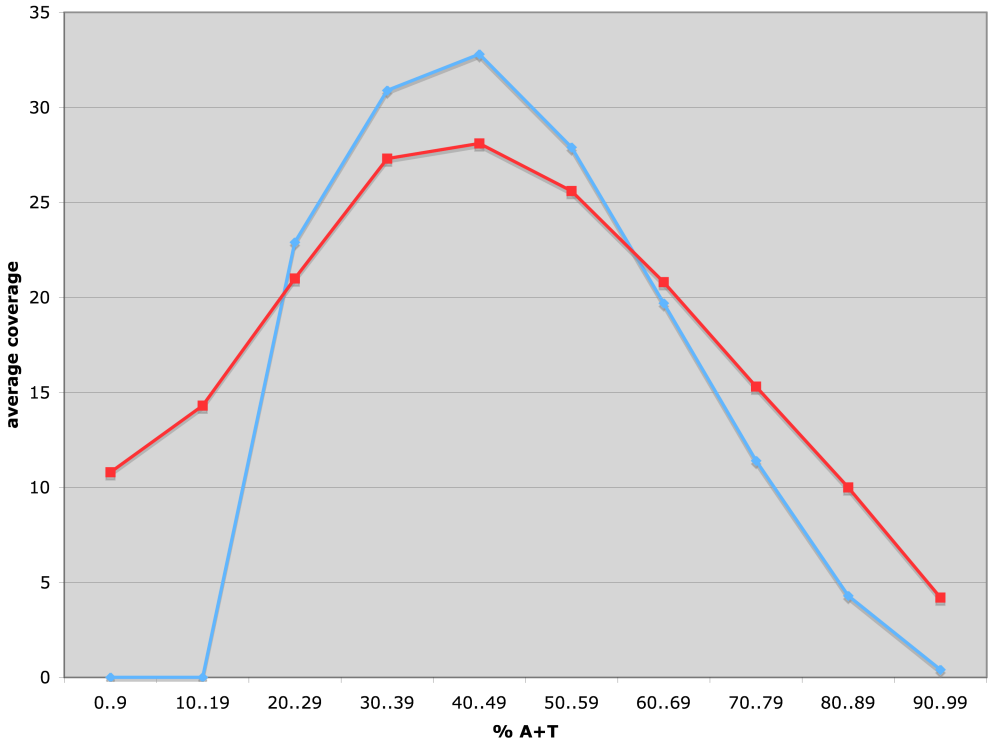
Supplementary Methods

Supplementary Table 1 Results of hairpin formation potential analysis in regions of the *C. elegans* genome with no exact match coverage by Solexa reads.

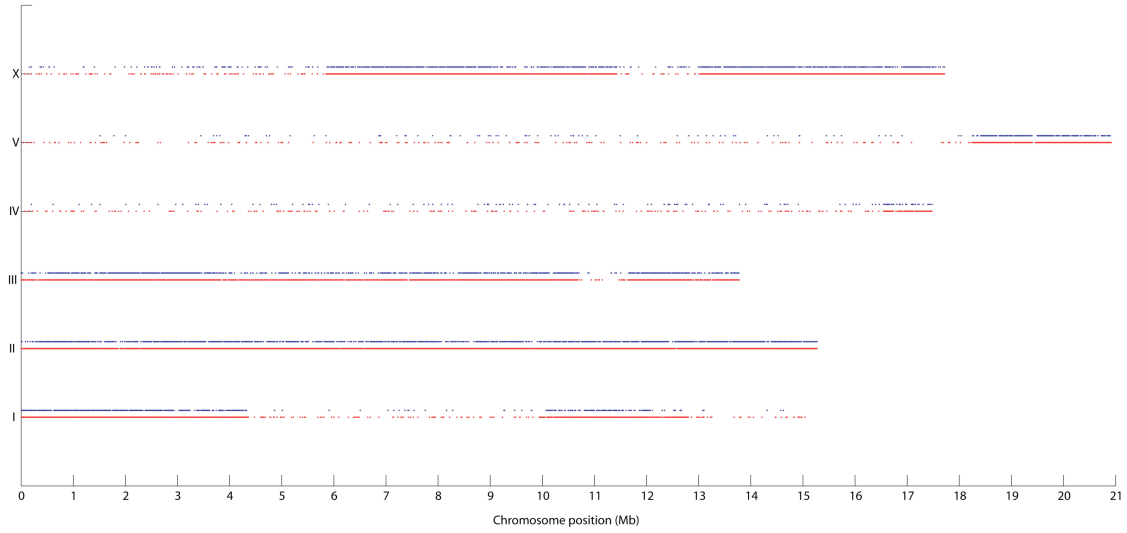
Supplementary Figure 1. Solexa single end read coverage levels mapped onto the C. elegans genome. Solexa single end read coverage of exactly matching reads mapped to the unique portion of the *C. elegans* reference genome. Average coverage was calculated at 19.2X.



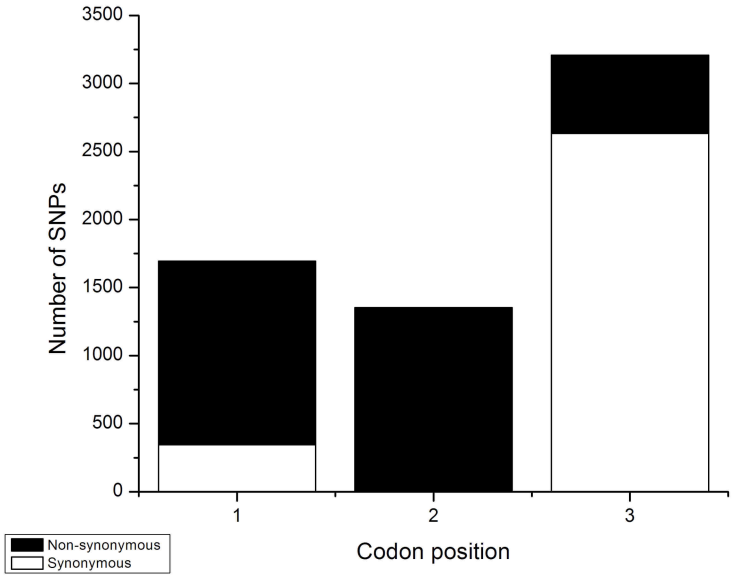
Supplementary Figure 2. Average N2 Bristol Solexa single end read coverage at different A+T percentages. The curves indicate the average coverage per base in A+T percentage bins ranging from 0-99 percent considering either a window size of 200 bp (red line), indicative of amplicon-associated biases or of 32 bp (blue line), indicative of read-associated biases.



Supplementary Figure 3. Chromosomal distribution pattern of CB4858 SNPs and indels. Chromosomal SNP (red) and indel (blue) distribution across the *C. elegans* genome, based on our analysis of CB4858 sequence alignment to the N2 Bristol reference genome sequence and variant detection by *PolyBayes*.



Supplementary Figure 4. Codon position bias in validated CB4858 SNPs. N2 Bristol-CB4858 codon position distribution of non-synonymous (black) and synonymous (white) SNPs, identified by analysis of CB3848 Solexa single end reads in comparison to the N2 Bristol reference sequence.



Supplementary Data

Resequencing of a *C. elegans* N2 Bristol strain isolate

Identifying unmapped read placements

Of our phrap-assembled contigs, 12,635 aligned with only a single mismatch to the *C. elegans* reference. An additional 3,088 contigs align with two mismatches, while 2,295 match over fewer than 50% of their bases. We further identified 124 contigs spanning one or more bases determined previously to be not covered (78 were gaps of 3 or fewer bases), suggesting insertion/deletion differences between the sequenced isolate and the reference. Thus, some unplaced reads simply are due to sequence differences between two isolates, not to missing sequences in the reference.

Evaluating Solexa single end read coverage

We calculated the ratio of observed to expected coverage by dividing the number of alignments that included each base by the number of times that base is involved in a 32mer in the *C. elegans* genome and found the average coverage ratio was 0.316, with the peak of the distribution around 0.25. This result is as expected, given that the peak of the coverage curve occurs between 16-17X (**Fig. 1**), and the representation value for a unique-in-genome base is 64, equaling a coverage ratio of 16/64=0.25. Paired end data provided a read-based average of expected to observed coverage of 0.014, with a corresponding physical coverage based average of 0.052.

To define a threshold for a significantly "over-represented" region (suggesting missing sequence from the reference or a possible coverage bias), we examined the Solexa read coverage in unique regions of the *C. elegans* genome (i.e. where the base had a representation value of 64 and was not in an annotated repeat in the *C. elegans* wb170 version at Wormbase). At the extremes, we found 1.32M base positions (1.7%) with >40X coverage in unique 32mers. Two areas of expected over-coverage are the rDNA regions on chromosomes I and V where uncertainty about the exact copy number of these large tandem repeats led to the placement of sole representative members of each in the reference sequence. Specifically, on chromosome I, the 18S, 28S and 5.8S are transcribed separately by RNA polymerase I in the ~55 copies of the 7.2kb rDNA repeat [1,9]. On chromosome V, the 5S gene along with the SL1 spliced leader gene lies in a 1kb tandem repeat with ~110 copies [9,10]. Examining unique 32mers within the rDNA segments, we ascertained that the unique chromosome I rDNA 32mers exceed a coverage ratio of 100 in Solexa reads; the only unique regions in the genome with this high a coverage level.

Analysis of regions not covered by Solexa reads

We further investigated the characteristics of regions not covered by Solexa reads, pursuing three hypotheses; 1) representational bias due to the potential for hairpin formation, 2) decreased coverage as a function of A+T content at the 32mer level and at the 200bp amplicon level (the fragment size utilized for Solexa libraries) and 3) correlation to sequence features such as repeats, exons, introns, etc. One possible explanation for uncovered regions, the formation of a hairpin that could cause deletion of the intervening sequence prior to Solexa library construction, was tested by examining the 31bp flanking each uncovered region for 5bp or more of sequence identity on

complementary strands. A control random set of flanking 31bp sequences was created for random pseudo-gaps across the genome (randomizing the chromosome and the start site and with the "gap" or "uncovered" size mimicking those of the true uncovered regions). We then required any sequences having 5 or more base pairs of identity to also have those identical sequences align immediately at the boundaries of the uncovered region. This analysis provided evidence of putative hairpins (**Table 1**) for the real and control gaps, and indicated that hairpin formation might account for only a handful of the uncovered genomic regions.

Our investigation of A+T content at the read- and amplicon-length level provided a compelling explanation for uncovered sequences. For zero base pair gaps, the A+T content of all 32mers centered on the base revealing the zero size gap was 76% and of all 200bp centered on that base was 73%. For gaps in coverage larger than zero bases, the AT content was 82%. When we stratified these regions by size, we found that regions between 25-100 bp had the highest A+T content; remarkably 85%. Examining regions covered by 2 or fewer reads and indicated an 84% A+T content.

To evaluate whether A+T biasing was occurring during cluster amplification or sequencing, we examined A+T content in 32 bp "read" windows compared to 200bp "amplicon" windows, then evaluated coverage of Solexa reads as a function of A+T content, examining the distribution for all bases and for all non-repetitive bases with a representation value = 64 (thus unique in genome). We considered these distributions for only 32mer exact match reads, and for the combined 32mer exact match and quality-trimmed exact match with 2 or fewer mismatches. For all permutations attempted we found the same trend; as A+T content reaches either extreme (very low or very high), the Solexa read coverage decreases. This is illustrated (**Fig. 2**), where a higher percentage of bases have zero coverage for a 200bp amplicon window, compared to a 32bp read window (51.6% vs. 9.7% for non-repetitive bases) suggesting that A+T content biasing may be related to the amplicon rather than the sequence read.

Finally, we correlated the locations of 13,035 regions without an exact match of >1 bp with other sequence features by number of events. Here, 33.4% of the uncovered base events occurred within repeats, 51.1% within introns, 2.3% within exons and 2.1% within "tagged" regions (annotated in the reference sequence as being a problematic region during conventional sequence finishing). Therefore, the relatively lower A+T content of *C. elegans* exons means that we can adequately characterize variant bases in exons, but this may not be the case for an organism with higher exonic A+T content.

Paired end coverage analysis

Although our paired end read coverage was significantly lower than from single end reads, it represented portions of the genome not represented in the other set, suggesting it may not have the same biases. Of the ~100kb of bases not covered by exact matching single end reads, 41% were covered by a read in the paired end set. Further, 80% of the not covered bases fell between paired ends (thus represented by physical but not by sequence coverage). The A+T content of the 49.5 Mb not covered by reads in the paired set, however, was 63%, similar to the genome average as a whole. Additional paired end coverage would be required for a definitive bias evaluation.

Genome-wide polymorphism discovery in *C. elegans* strain CB4858

Mosaik alignment of CB4858 reads

Of the total 37,856,444 CB4858 Solexa reads, the Mosaik aligner/assembler was able to align 24,518,452 (64.7%) to the conservatively masked *C. elegans* reference genome (ws170), allowing no more than two sequence differences (mismatches plus gaps), to avoid misalignments due to paralogy.

Chromosomal locations of CB4858 polymorphisms

In 2000, Koch[1] illustrated in *C. elegans* that the non-synonymous substitution rate was much higher in the first and second codon positions than in the third. Our study confirms these earlier results (**Fig. 3**) and provides a detailed, genome-wide estimate of coding polymorphisms in the CB4858 strain. In total, we found 6,255 SNPs positioned within an exon, of which 3,275 putatively introduce an amino acid change. Through our experimental validation, 100 of 119 (84%) non-synonymous mutations were confirmed, indicating that our methods provide an important first step in describing the complete mutational profile in a strain-to-reference paradigm. Furthermore, we evaluated SNP positioning on a chromosome-by-chromosome basis (**Fig. 4**), finding the resulting polymorphism density to be much higher on Chromosomes II, III and X. The densities on Chromosomes IV and V suggest a very low variation rate for much of the chromosome, yet a comparatively high mutation rate on the right half of each chromosome.

Supplementary Methods

Preparation of Solexa fragment libraries

Genomic DNA (5μg) was nebulized for 2 minutes at 45 psi of compressed air, to obtain an average fragment size of 500bp, then further purified and concentrated with Qiaquick PCR purification spin columns (Qiagen Inc., Valencia CA). Treatment to remove 3' overhangs and fill in 5' overhangs resulted in blunt ended genomic fragments. An A residue was added by terminal transferase to the 3' end and the resulting fragments were ligated with Solexa adapters. Adapter-modified DNA fragments were enriched by an 18 cycle PCR using 50 ng of the ligation reaction and Solexa universal adapter primers. The resulting PCR products were separated by agarose gel electrophoresis and the band between 150-200 bp was excised from the gel. The DNA fragments were extracted from the agarose slice using a Qiaquick Gel Extraction Kit (Qiagen Inc.), and further purified by drop dialysis using a 0.025mm/25mm filter (Millipore Inc., Billerica MA) and tissue culture-grade water (Sigma Chemical, St. Louis MO). The DNA fragment library was quantitated, then diluted to a 10 nM working stock for cluster generation.

Solexa cluster generation

Adapter-ligated fragments (2 nM) were denatured in 0.1N NaOH for 5 minutes, then were further diluted to a final 9 pM concentration in 1 ml of pre-chilled hybridization buffer, and introduced onto the Solexa flow cell using the Cluster Station, an automated device supplied by Solexa. On this apparatus, the oligo-derivatized flow cell surfaces hybridize to library fragments by adapter-to-oligo pairing. "Clusters",

representing discrete populations of unique single-stranded library fragments amplified *in situ*, are generated by isothermal amplification using a proprietary process. In practice, each cluster produces a single Solexa read. Our experiments aimed for an average cluster density of 30,000 (+/- 5,000) clusters per flow cell lane.

Preparing clusters for sequencing

Following isothermal amplification, clusters were made single-stranded by 0.1N NaOH denaturation, metered across the flow cell by the Solexa Cluster Station. A sequencing primer complementary to one Solexa adapter was added to prime the single-strands of each cluster. Once hybridized and with excess primer removed by a wash, the flow cell was ready for sequencing.

Solexa single end read sequencing process

The Solexa Genome Analyzer was programmed to provide up to 32 sequential flows of fluorescently labeled, 3'-OH blocked nucleotides and polymerase to the surface of the flow cell, thus producing a fixed 32bp read length. After each base incorporation step, the flow cell surface is washed to remove reactants and then imaged by microscope objective. Our experiments collected 200 tiled images ("tiles") per flow cell lane, each containing on average 30,000 clusters. Solexa single end reads were generated for N2 Bristol (85,498,849 total reads) and CB4858 strains (37,856,444 total reads) using a 30 base read length for the titration run (used to determine the correct input library DNA amount), and a 32 base read length for standard runs.

Solexa paired end N2 Bristol sequencing process

Paired end library construction involved the DNA fragmentation steps outlined above, but ligated the end-repaired fragments with Solexa paired end adapters to introduce two unique sequence priming sites at each fragment end. The adapter-ligated fragments were separated by agarose gel electrophoresis and a scalpel was used to cut a thin horizontal slit in the lane at 200bp. The DNA fragments from this slit were rinsed from the scalpel blade using 30 μl of EB (Qiagen Inc., Valencia CA), then were enriched by an 18 cycle PCR (conditions described above). The resulting PCR products were purified and concentrated with a Qiaquick PCR purification spin column (Qiagen Inc.). The paired end library was quantitated and diluted to a 10 nM working stock. Cluster generation was accomplished on a modified flow cell that contained three primer-complementary sequences. These sequences promote the formation of heterogeneous clusters in which two populations exist; one that can be subsequently released for sequencing primer annealing using the standard chemistry, and a second population that remains intact during release and sequencing of the first population, then is subsequently released for sequencing primer hybridization and extension by a separate chemical step that is applied whilst the flow cell remains in place on the Solexa Genome Analyzer. A total of 3,671,972 paired end reads were produced, with a first end read length of 35 bp and a second end read length of 25 bp.

Solexa data processing pipeline

The raw images from the Solexa Genome Analyzer are synchronously mirrored as the run proceeds, from the instrument computer to an NAS (Network Attached Storage)

filesystem. Once the run completes, image analysis, base calling and read quality filtering are performed by the Solexa run analysis software. The read quality filtering acts to first remove poor quality reads using Solexa-defined criteria ("% passed bases"), then uses the program Gerald to align and compute the percentage of bases aligning to the reference *C. elegans* genome (ws170 version). The alignment of the resulting short reads to the reference genome is performed by the Solexa pipeline. Although we did not utilize the Eland alignments in the detailed data analyses described here, Eland metrics did provide an initial evaluation of Solexa data quality and helped us to estimate accumulating coverage.

Solexa N2 Bristol reads-to-reference alignment approach

We initiated read alignment to the reference by first identifying all exactly matching reads. Here, we generated a hash table of all 30mers and 32mers (to match the Solexa read lengths on titration and full runs) in the reference *C. elegans* genome sequence, and the positions for each mer within the *C. elegans* genome were cataloged. The UNIX sort function was then used in a hash-based approach, to sort the Solexa reads with all reference-derived 32mers and 30mers, determining that 75.55% of reads had an exact match. Reads were placed in all regions where they had an exact match. Of the reads without an exact match, we removed all reads with a best match (by BLAT) to *Escherichia coli* (*C. elegans* feed on *E. coli*). Further, we removed all reads (1.05%) matching a portion of the Solexa primer sequence (which has no match in either *C. elegans* or *E. coli*).

The remaining reads were quality trimmed, meaning we retained reads with at least 20 consecutive base pairs having a quality score of 25 or higher (36.3% retained). These reads were again sorted by exact match hash-based methods to an index of all 20mer, 21mers, etc. through 32mers for the *C. elegans* and *E. coli* reference sequences. The remaining quality-trimmed reads without an exact match to *C. elegans* or to *E. coli* were then considered as potentially falling into one of the following categories: 1) contamination, 2) containing one or more base calling errors, 3) indicative of potential errors in the reference genome or of a difference between the reference and Solexa-sequenced strain, 4) sequences altogether missing from the current reference genome.

We used BLAT-based[14] alignments with the parameters (tileSize=6 –stepsize=1 repMatch=1000000 –minIdentity=70 minScore=20) to identify non-exact matching reads. These were identified as Solexa reads placed uniquely by gapped alignment, that indicated an insertion in the Solexa-sequenced N2 strain relative to the reference (defined by an exact Solexa read alignment on either side of a one base (or larger) gap), or uniquely-placed Solexa reads with a deletion in the Solexa-sequenced strain (defined by exact alignment that excluded one or more unmatched bases in the reference sequence). We also found evidence of indels, where the Solexa read and the reference matched exactly except for one or more bases that were not matched between the two in the unaligned region.

Assembly and analysis of unplaced reads

To investigate the remaining 1,843,353 unmapped reads, we attempted assembly using phrap (P. Green, unpublished) with stringent parameters (-minmatch 20 –minscore 20 –forcelevel 0 –repeat_stringency 1 –new_ace_penalty -4). Only 331,568 reads

assembled into 34,382 contigs with an average length of 35 bp. These contigs were compared to the reference genome using BLAT. The remaining reads likely differ significantly or are missing completely from the *C. elegans* or *E. coli* reference sequences (*E. coli* strain OP50 is fed to *C. elegans*), or are contamination or sequencing artifacts.

Solexa N2 Bristol single end read coverage

To identify regions with above- or below-average coverage by Solexa N2 Bristol reads, and to evaluate the potential for biases in coverage, we first created a map enumerating Solexa reads in which each basepair in the *C. elegans* genome participates. If a base is in a unique region of the genome, the "representation value" for that base is 64 since it will only be located in the 32mer including that base, as well as on the opposite strand. Of the 100,281,244 bases in the *C. elegans* reference, 87,558,992 bases participate in a single 32mer (i.e. representation value of 64). To obtain an estimate of unique genome coverage, we then examined the number of times each 32mer occurred within the reference genome.

Evaluating sequences not covered in the reference

Genomic regions that remained not covered by Solexa reads were identified as all bases not participating in an alignment with a Solexa read after mapping all exactly matching reads and quality-trimmed reads, and by further mapping quality-trimmed reads allowing up to 2 mismatches or unaligned bases. In addition to obvious large gaps in coverage, our gap analysis also revealed so-called "zero base pair gaps". We defined a zero base pair gap as a site where two adjacent reference genome 32mers are covered by reads that align exactly to each 32mer but across which no exactly matching spanning read exists. All of the uncovered regions we computed were evaluated for their A+T content, potential for hairpin formation, presence of transposon sequences, and other sequence features.

Paired end alignment

Of the 3,671,972 paired end reads produced, 2,158,677 (58.8%) forward reads and 2,663,404 (72.5%) reverse reads had an exact match. The difference in matching percentage likely was due the lower base calling accuracy near the end of the longer forward read. We identified 1,581,763 N2 Bristol quality filtered read pairs that could be uniquely placed in the genome and had an average insert size of 218 bp (s.d. = 38 bp). These were evaluated further for their potential to identify putative areas of structural variation, as described in the Results section of the manuscript.

Mosaik alignment of CB4858 Solexa reads

Our approach used two fundamental methods: (1) hash-based, and (2) sequence alignment-based. Our hash-based method enumerated every 32mer in the *C. elegans* reference genome sequence, and recorded its map location. If the same 32mer occurred in multiple locations (either strand), it was marked as a microrepeat. Near-perfect microrepeats were identified by asking whether each specific 32mer, or any of the other 32mers obtained by introducing "mutations" up to a pre-specified number of mismatches, appeared at any other genomic location. The sequence alignment-based method used the BLAT algorithm with the following parameters: -stepSize=8 -tileSize=16 -minMatch=1 -

minScore=28 -oneOff=1 to enumerate every 32mer in the reference genome, and then to search the rest of the genome for a perfect or near-perfect match, up to a pre-specified number of mismatches. Custom scripts identified hits that had up to 2 mismatches (any combination of substitutions, insertions, or deletions), and combined the results of our hash-based and sequence alignment-based methods to produce a microrepeat-masked reference genome.

Mosaik consists of two parts: the aligner and the assembler. The aligner employs a two-stage process: an initial short hash-based search to explore possible map locations for each read and to register locations with multiple initial hash matches, followed by a full Smith-Waterman-Gotoh pair-wise local alignment between the read and each possible map location. The best-scoring map location is identified, and the alignment is reported if 1) the alignment score is above a pre-specified threshold, 2) the alignment extends over a pre-specified fraction of the reads, and 3) the total number of mismatches within the alignment does not exceed a pre-specified maximum. In this analysis, we allowed a maximum of two mismatches (substitutions and indels), a decision motivated by the fact that over 79% of the Solexa reads had either one or zero error. Allowing a single polymorphic difference in such reads would bring the maximum number of mismatches to two. The assembler algorithm first registers every gap introduced by any of the aligned reads on the reference sequence, then introduces alignment gaps into all aligned reads to preserve the positions of all pair-wise aligned bases in every pair-wise read alignment.

Validation of putative sequence differences

A subset of candidate insertion, deletion, indel and polymorphic sites from our analyses were submitted for orthologous validation using PCR-based sequencing and variant analysis. Because we had a very limited amount of the original N2 Bristol genomic DNA isolate used for the reference genome sequencing, we had a similarly limited ability to validate the sequence variants we identified between Solexa N2 Bristol reads and the reference genome. Hence, the preponderance of our validations centered on the putative variants identified for the CB4858-to-N2 Bristol comparison.

For each type of CB4858 variant, we attempted validation for 50% of the candidates identified in coding and 50% in non-coding sequences. We designed primers with 300bp of sequence both to the left and to the right of the target site, using Primer3 (http://primer3.sourceforge.net/). A 5' universal M13 forward or reverse sequence was added to each primer pair to allow processing of the resulting PCR products in our high-throughput sequencing pipeline. PCR was performed in a 10μL reaction containing 5ng of genomic DNA (the original DNA stock utilized in Solexa sequencing, or the original N2 Bristol reference strain DNA, according to the assay type), 1.2nmol of each universally tailed amplification primer, and 5.0μL of Amplitaq Gold 2X mix (Applied Biosystems P/N 4327059, Foster City,CA), with a final 8% glycerol concentration). Reactions were cycled at 96 ºC for 5 minutes, followed by 40 cycles of 94 ºC for 30 seconds, 60 ºC for 45 seconds, and 72 ºC for 45 seconds, and a final extension at 72 ºC for 10 minutes. Following amplification, PCR products were treated with 3.7U of Exonuclease I (USB P/N 70073X, Cleveland,OH) and 0.18U of Shrimp Alkaline Phosphatase (USB P/N 70092X), and incubated at 37 ºC for 30 minutes. The reaction was stopped by incubation at 80 ºC for 15 minutes. PCR products were sequenced with

BigDye Terminator Sequencing Kit (Applied Biosystems P/N 4336943, Foster City CA) using either universal forward or reverse sequencing primers, and analyzed on ABI 3730XL DNA sequencers.

Visual inspection of validation data

We aligned validation reads to the reference sequence using PolyPhred[2], determining by manual inspection whether a Solexa variant was confirmed in the 3730 trace. Both a validation rate (defined as the number of confirmed variants divided by the number of successful sequencing reactions) and a conversion rate (defined as the number of confirmed variants divided by the number of variants submitted for validation) were calculated based on our visual confirmation results.

Evaluating variants for exonic disruption

We investigated SNP candidates in CB4858 that introduced amino acid changes relative to N2 Bristol. Conceivably, such differences might suggest subtle chemosensory or other adaptations specific to CB4858. Using the Wormbase gene annotations, we characterized all SNPs that were within exons by their codon position and whether the variant caused a synonymous or non-synonymous amino acid change.

Supplementary Table 1. Results of hairpin formation potential analysis in regions of the *C. elegans* genome with no exact match coverage by Solexa reads. The random set data were generated as a control to gauge the significance of any biases indicated by evaluation of the real set.

Basepairs of flanking sequence	Number of putative hairpins in real set	Number of putative hairpins in random set
5	47	46
6	34	19
7	27	11
8	23	5
9	18	4
10	13	1
11	11	1
12	9	1
13	8	1
14	8	1
15	6	1
16	5	1
17	5	1
18	5	1
19	5	1
20	5	1
21	5	1
22	5	1
23	5	0
24	4	0
25	4	0
26	4	0
27	4	0
28	3	0
29	1	0
30	0	0
31	0	0

References

[1] R. Koch, H. G. van Luenen, M. van der Horst et al., *Genome research* **10** (11), 1690 (2000).

[2] D. A. Nickerson, N. Kolker, S. L. Taylor et al., *Methods in molecular biology (Clifton, N.J* **175**, 29 (2001).