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

LaDeana W Hillier[1,3], Gabor T Marth[2,3], Aaron R Quinlan[2], David Dooling[1], Ginger Fewell[1], Derek Barnett[2], Paul Fox[1], Jarret I Glasscock[1], Matthew Hickenbotham[1], Weichun Huang[2], Vincent J Magrini[1], Ryan J Richt[1], Sacha N Sander[1], Donald A Stewart[2], Michael Stromberg[2], Eric F Tsung[2], Todd Wylie[1], Tim Schedl[1], Richard K Wilson[1] & Elaine R Mardis[1]

Massively parallel sequencing instruments enable rapid and inexpensive DNA sequence data production. Because these instruments are new, their data require characterization with respect to accuracy and utility. To address this, we sequenced a *Caernohabditis elegans* N2 Bristol strain isolate using the Solexa Sequence Analyzer, and compared the reads to the reference genome to characterize the data and to evaluate coverage and representation. Massively parallel sequencing facilitates strain-to-reference comparison for genome-wide sequence variant discovery. Owing to the short-read-length sequences produced, we developed a revised approach to determine the regions of the genome to which short reads could be uniquely mapped. We then aligned Solexa reads from *C. elegans* strain CB4858 to the reference, and screened for single-nucleotide polymorphisms (SNPs) and small indels. This study demonstrates the utility of massively parallel short read sequencing for whole genome resequencing and for accurate discovery of genome-wide polymorphisms.

In 1998 the decoding of the first animal genome sequence, that of *C. elegans*, was published[1]. *C. elegans* was first suggested as a model organism in the 1960s by Sydney Brenner, and subsequent work produced a physical map of its genome[2]. As a result, the *C. elegans* genome sequencing project formed the cornerstone of efforts ultimately aimed at decoding the human genome[3,4]. The entire *C. elegans* biology community has benefited enormously from the availability of the genome sequence and the ever-improving genome annotation[5], and from the comparative power of the availability of sequenced genomes for *C. elegans*' relatives such as *C. briggsae*[6].

The emerging availability of massively parallel sequencing instrumentation provides the capability to resequence genomes in a fraction of the time, effort and expense than ever before. Compared to capillary sequencing, these instruments produce relatively short-read-length sequences that require characterization, including read error profiles and base call accuracy (which we refer to as base quality) values. Furthermore, the general utility of short read sequences, coverage models for resequencing and approaches for read mapping to reference genomes requires investigation. To address these, we sequenced an isolate of the *C. elegans* N2 Bristol strain using the Solexa Sequence Analyzer (Illumina Inc.). Our analyses of these sequences included (i) an elucidation of the Solexa read error model, (ii) an evaluation of sequence differences between the two isolates and (iii) identification and investigation of representational biases in Solexa data. We revealed possible sequencing errors in the *C. elegans* reference genome, and putative variants that had occurred in our passaged N2 Bristol strain.

Massively parallel sequencing can be applied to strain-to-reference comparisons that reveal genome-wide sequence differences, either for evolutionary studies or for discovering genetic variation that may explain phenotypic variation. Implementing this application requires a new approach that assesses the fraction of a genome to which short read sequences can be uniquely mapped because they are more susceptible to multiple placements than are longer capillary instrument–derived sequences. Computational identification and markup of these 'microrepeats' is therefore an important precursor to accurate short-read analysis, and must allow for mismatches resulting from sequencing errors or polymorphisms. We aligned Solexa sequence reads from the *C. elegans* strain CB4858 (originally isolated in Pasadena, California, USA)[7] to the microrepeat masked N2 Bristol reference sequence, and identified SNPs and small indels with a modified PolyBayes[8] version. Orthologous validation yielded a high validation rate.

RESULTS

Experimental design

In this study we explored two applications of Solexa sequencing: (i) genome resequencing and (ii) genome-wide polymorphism discovery. For the first application, we sequenced an isolate of the *C. elegans* N2 Bristol strain at a high coverage depth with single-end reads and at a much lower coverage depth with paired-end reads. Using the reference genome as our alignment target, we determined

[1]Washington University School of Medicine, Department of Genetics and Genome Sequencing Center, 4444 Forest Park Blvd., St. Louis, Missouri 63108, USA. [2]Boston College, Department of Biology, 140 Commonwealth Ave., Chestnut Hill, Massachusetts 02467, USA. [3]These authors contributed equally to this work. Correspondence should be addressed to E.R.M. (emardis@watson.wustl.edu).

RECEIVED 19 SEPTEMBER 2007; ACCEPTED 21 DECEMBER 2007; PUBLISHED ONLINE 20 JANUARY 2008; DOI:10.1038/NMETH.1179

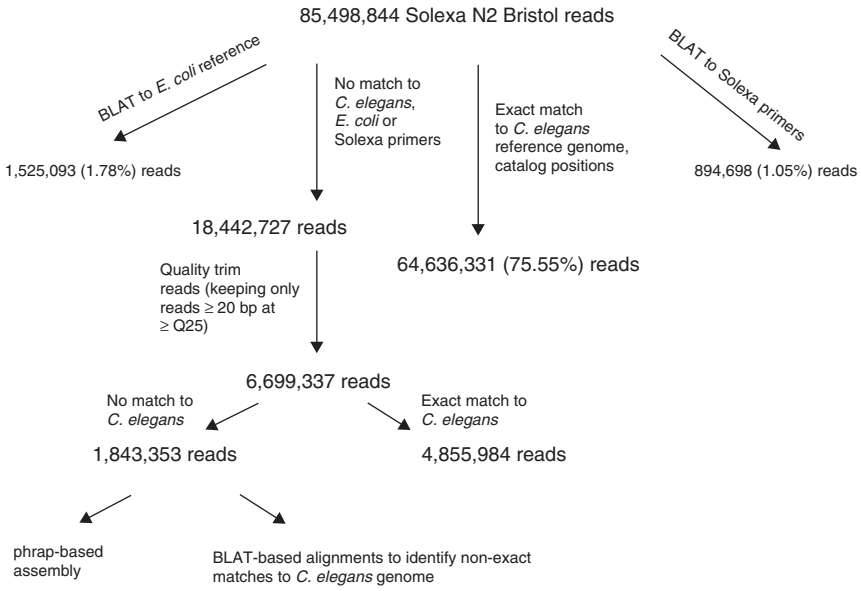

Figure 1 | N2 Bristol Solexa read analysis. The diagram shows the processing steps used to evaluate Solexa single-end reads from the N2 Bristol isolate. The majority of reads mapped exactly to the reference genome.

an accuracy estimate and an error model for Solexa reads. We next aligned all reads possible using a tiered approach (**Fig. 1**), identified sequence differences between the two isolates and evaluated both representational bias and copy-number detection.

We developed a genome-wide polymorphism discovery approach by first sequencing *C. elegans* strain CB4858, using Solexa single-end reads of about ninefold coverage. To decrease the possibility of erroneous variant detection because of paralogous read placements, we identified and masked 'microrepeat' regions in the genome based on a 32-bp read length. We then aligned CB4858 reads to the reference genome using Mosaik and applied a modified PolyBayes version to detect variants. Our predicted polymorphic sites were validated by PCR amplification and Sanger sequencing at a high rate.

Resequencing a *C. elegans* N2 Bristol strain isolate

We used the single-end *C. elegans* N2 Bristol reads to evaluate the overall accuracy and quality of Solexa pipeline passed reads. **Table 1** provides several metrics of our Solexa single-end read dataset, including Eland alignment results to the ws170 release of the *C. elegans* reference genome[9] (http://www.wormbase.org). Based on the Eland metrics, we estimated 20-fold coverage of N2 Bristol for the quality passed and aligned single-end reads.

We performed read alignment with EagleDiscoverer, and subsequent error analysis revealed that 57.2% of the uniquely mapping single-end reads contained zero mismatches and 79.9% had 0 or 1 mismatch. We determined the full distribution of mismatches for the Solexa N2 Bristol reads (**Fig. 2**), and the position-specific dependency of Solexa base calls at phred qualities of 25 and 30 (**Fig. 3**), which illustrates that the base accuracy for a given base quality value depends upon that base's position in the sequence read.

To determine Solexa N2 Bristol single-end read coverage, we first devised an iterative read-alignment strategy for these reads (**Fig. 1** and **Supplementary Methods** online). We then determined the average coverage of the genome to be 19.2 (s.d. = 9.0; **Supplementary Fig. 1** online).

We examined the genome for over-represented regions and found ~1.7% of the genome had >40-fold coverage in unique 32-mers. We expected ribosomal DNA genes to have higher than average coverage because the reference represents these as single copies but they actually exist in multiple copies. By examining unique 32-mers within rDNA segments, we found evidence of excess coverage (>100× for the chromosome 1 rDNA unique 32-mers). This finding lends credibility to the use of read coverage as a quantitative metric of region-specific copy number. Based on our analysis of regions with higher than average coverage, combined with the assembly and analysis of unmapped reads (see **Supplementary Methods**), we estimate a maximum of ~0.5 Mb of repetitive sequence is missing from the *C. elegans* reference genome.

To determine the genome coverage of the sequence reads we proceeded as follows. After aligning exactly matching 32-bp reads, exactly matching quality-trimmed reads (that is, reads with at least 20 consecutive base pairs of quality score ≥20) and quality-trimmed reads with 2 or more mismatches to the reference genome, we found that 99.9% of the unique *C. elegans* genome was covered, mostly in large spans; the longest was 194 kb on chromosome V. Of the regions left uncovered by Solexa reads, there were 9,492 gaps comprising 95,913 bp. These coverage gaps ranged in size from 1 base to over 1,000 bases; 77.9% were 1–9 bp and another 36.8% were 10–50 bp. If we only consider 32-mer exact mapping reads, the largest coverage gap was a 4,601-bp region on chromosome X (5907815–5912415). Notably, the entire 4,600-bp region is bounded by a transposon (TC5A#DNA/Tc4) in the reference sequence and is completely contained within a single fosmid clone (H05L03) that extends into an overlapping yeast artificial chromosome (YAC; Y23B4A).

There were a total of 1,728 zero-base-pair gaps discovered—areas of the genome where two adjacent reference genome 32-mers were covered by reads that aligned exactly to each 32-mer but across

Table 1 | Solexa run metrics for N2 Bristol and CB 4858 single-end reads

Genome	Number of runs	Total bases generated	Total passed bases	Percentage passed bases	Percentage aligning (ws170)	Percentage error (alignment-based)
N2 Bristol	3.5	4.06 Gb	2.67 Gb	66%	79%	0.6%
CB4858	1.5	2.52 Gb	1.35 Gb	54%	67%	0.52%

Solexa run metrics obtained for the combined 30 and 32 bp single-end reads from both the N2 Bristol and CB 4858 isolates. Results, including the total number of bases generated, the total number of passed (for example, high quality) bases, and the percentage of aligning reads were obtained from the output of the Solexa-provided data analysis pipeline. The Eland-generated error rate is reported, based on the reference genome alignments of Solexa passed reads.

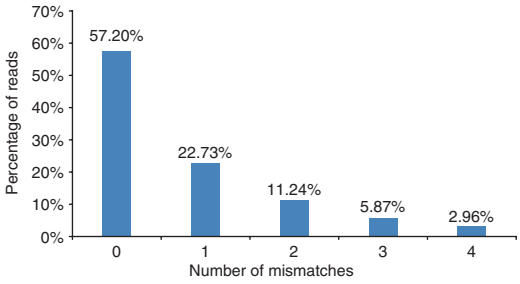

Figure 2 | Accuracy distribution of N2 Bristol Solexa single-end reads. As described in the text, after alignment of N2 Bristol Solexa reads to the reference genome sequence using EagleDiscoverer and tabulating any differences between the two sequences, we determined that ~80% of the reads exhibited 0 or 1 mismatch when uniquely aligned to the reference genome.

which no exactly matching spanning read exists. Of these, 1,564 (90%) had single read representation of the flanking 32-mer, consistent with the notion that these regions are under-represented by Solexa reads. Further investigation of these coverage gaps revealed that (i) they are located primarily in noncoding sequence (2% are in exons), (ii) only a few regions could be explained by hairpin formation (see **Supplementary Data** and **Supplementary Table 1** online) and (iii) the A+T content in these regions is substantially higher than the genome average (85% versus 65%, respectively; **Supplementary Fig. 2** online). Furthermore, this A+T bias more likely occurs during amplification than during sequencing (see **Supplementary Data**). We identified 125 zero-base-pair gaps with a non-identical spanning Solexa read, suggesting an insertion in the Solexa-sequenced strain. We resequenced 22 of these and validated them as true differences between the two N2 Bristol isolates.

Genomic alignment of nonexact match reads (that is, N2 Bristol single-end reads without an exact match to *C. elegans*; **Supplementary Methods**) allowed us determine differences between the two N2 Bristol isolates and to identify possible errors in the reference sequence as these reads are highly similar but contain inserted or deleted bases that preclude an exact match. Here three or more Solexa reads were required to predict a reference error, to reduce contributions from Solexa base calling errors. Such alignments putatively identified 2,981 insertions, deletions and indels of 1–20 bp. Of these, 2,082 occurred at positions also having exactly matching Solexa reads, thus confirming the reference sequence and indicating an allelic polymorphism between the two N2 Bristol isolates. By contrast, 235 of the indels occurred in regions with no perfectly aligning Solexa read, suggesting a possible error in the *C. elegans* reference genome, and indicating a potential indel error rate of 1 in 373 kb. We detected 56 different putative deletion events, for which a Solexa read spanned one or more bases in the reference genome, aligning immediately on either side. Alternatively, these could be insertions in the reference genome. Lastly, 53 different putative insertions were suggested (by 502 reads) for which a Solexa read had extra bases relative to the reference. These also could be deletions in the reference.

We identified 1,396 nonrepetitive, uncovered regions with at least one read having an unaligned or mismatched base, suggesting a Solexa base-calling error, a polymorphism in the Solexa-sequenced N2 Bristol isolate or a substitution error in the reference genome. Of these, 1,011 were covered by more than one read, and 544 were covered by more than two reads. These suggest a maximum substitution error rate in the *C. elegans* reference sequence of 1 in 99 kb. We included a limited number of these putative errors in our validation efforts, described below.

We were able to produce and analyze limited numbers of paired-end reads for *C. elegans* N2 Bristol, providing an average coverage of 0.84× with a mean physical coverage (measured by the span of matching paired ends) of 3.08×. Because paired-end reads are used to evaluate structural variation based on deviations in end read distance from expectation[10], we determined that 37,352 read pairs had a mapped distance >3 s.d. from the 218-bp average (**Supplementary Data**; 36,209 were <104 bp and 1,143 were >332 bp). If we required more than one read pair placement to confirm an event, only 5,908 pairs remained (5,670 were <104 bp). Notably, these 5,670 read pairs spaced ~100 bp closer than expected support our estimate that ~0.5 Mb of the *C. elegans* repetitive genome is missing from the reference genome. The vast majority of multiple read pairs that confirmed a structural variation event were within introns and/or were annotated as repetitive. Many such read pairs confirmed regions already annotated as "difficult to sequence"; about 1.5 times as many fell in genomic regions sequenced from YACs or plasmids (both clone types were used to sequence regions unclonable in cosmids for the reference genome sequencing). For example, on chromosome III, a complex tandem repeat annotated as "restriction digest data indicate 3 kb is missing from the assembly of this region" was identified by 238 Solexa paired ends placed at >3 s.d. apart (50% were 332–400 base pairs apart), further substantiating the initial suspicion of misassembly. Hence, paired-end data enhance the utility of Solexa reads, providing an important tool for identifying putative structural variation.

Polymorphism discovery in *C. elegans* strain CB4858

We sequenced an isolate of the CB4858 strain using the Solexa technology to produce ~ninefold coverage in single-end reads. This strain was selected because previous work had suggested

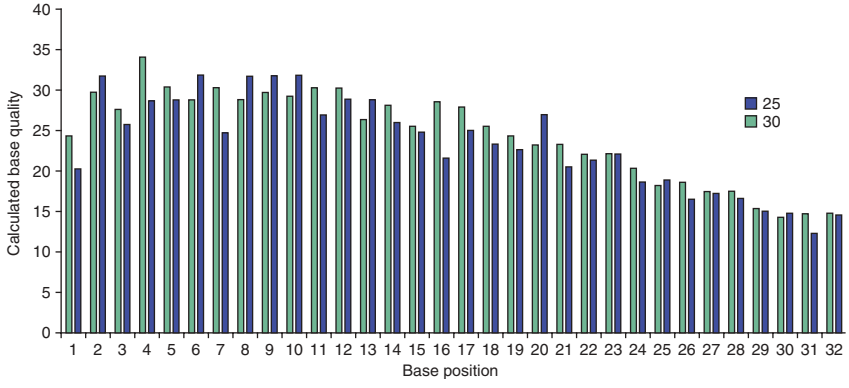

Figure 3 | Position dependency of base calling accuracy for N2 Bristol Solexa single-end reads. The calculated based quality is shown as a distribution at phred base quality 25 or at base quality 30.

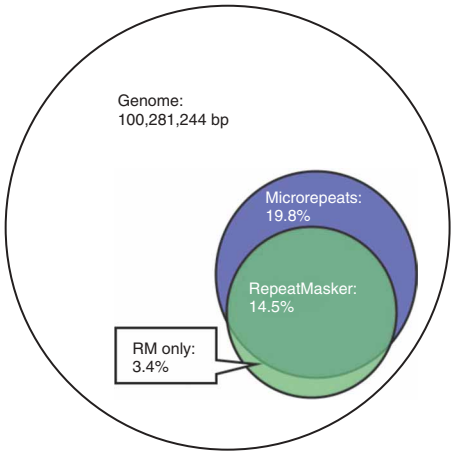

Figure 4 | Repetitive content in *C. elegans*. Venn diagram depicting the fraction of bases in the genome covered by microrepeats and by RepeatMasker, and the overlapping set.

a polymorphism rate of 1:1,600 (ref. 11). The Solexa analysis pipeline produced metrics of our Solexa single-end read data for CB4858 (**Table 1**).

As a precursor to variant discovery in CB4858, we identified regions of the reference genome with a high potential for ambiguous read alignment, based on the Solexa 32-bp read length. First, we identified all unique 32-mers in the reference sequence, but as our error rate analysis (**Fig. 2**) indicated a drop-off in the error rate beyond 2 errors per read, we defined a repetitive 32-mer as one that appears in the genome more than once, allowing 0–2 mismatched bases (substitutions, insertions or deletions). We called these 'microrepeats' to distinguish them from repeats marked by the RepeatMasker program[12], which masks 14.5% of the bases in the genome. The fraction of the genome comprising perfect and near-perfect microrepeats totaled 19.8%. We illustrate the relationship between RepeatMasker-masked bases and microrepeat bases identified by our methods as a Venn diagram (**Fig. 4**). Although there is a substantial overlap (11.11%) between the regions masked by both methods, 8.7% of the genome that we identified as microrepeats was not masked by RepeatMasker. Conversely, 3.4% of the genome was masked by RepeatMasker only, indicating that some fraction of *C. elegans* repeat elements can be uniquely sequenced with 32-bp reads. Taken together, RepeatMasker repeats and microrepeats cover 23.2% the genome.

Once we aligned CB4858 Solexa reads to the conservatively masked *C. elegans* genome, we applied our combined repeat masking to filter the alignments, identified high-quality sequence differences with PolyBayes, and finalized a set of 45,539 SNPs and 7,353 single-base-pair indels. This yields a rate of one SNP per 1,629.81 bp and one indel per 9,894.99 bp. Hence the pair-wise nucleotide diversity (theta) between the CB4858 and the N2 Bristol strains is 6.136×10^{-4}, in good agreement with the ~1:1,500 rate posited in a previous description of CB4858 (ref. 11). As 37,856,444 CB4858 Solexa reads yielded a total number of 45,539 SNPs, the 'read-per-SNP' yield was 831. All confirmed CB4858 sequence variants are available in Wormbase.

We orthologously validated roughly 1,000 candidate SNPs and indels by PCR-directed capillary sequencing to gauge the performance of our Mosaik-PolyBayes approach. After sequencing and evaluation, we determined a SNP validation rate of 96.3% (438/455) and an 89.0% conversion rate (438/492) for candidates identified by PolyBayes (**Table 2**). We sequenced 239 of our putative single-base indels, finding they validated (93.8%) and converted (87.7%) at practically the same rates as SNPs. Both insertions and deletions predicted in the reference genome sequence were represented (insertions: 2,948 or 47.1%, and deletions: 3,316 or 52.9%). Many of the indels were variable numbers of bases in mono-nucleotide repeats, for example, 5 versus 4 adenosines. Although mononucleotide runs are typically very difficult areas for indel detection, our high validation rate indicates that Solexa reads resolve base numbers in these runs very well. Microrepeat masking has a marked impact on accurate SNP discovery by eliminating putative SNPs and indels resulting from paralogous read mapping (**Table 2**).

We estimated false negative rates for PolyBayes by running PolyPhred[13–15] (version 5.0) on the validation trace data. This algorithm indicated PolyBayes had missed 26 SNPs, for a false negative rate of 3.75%.

To determine the chromosomal distribution of CB4858 polymorphisms, we placed CB4858 SNPs and indels along the six *C. elegans* chromosomes, and identified both chromosome-wide and chromosomal position–specific differences (**Supplementary Data** and **Supplementary Fig. 3** online). Our data confirmed an earlier study in *C. elegans*[16] suggesting that nonsynonymous substitution rates are higher in the first and second codon positions than in the third (**Supplementary Fig. 4** online). Furthermore, over half of CB4858 SNPs positioned in exons putatively introduce an amino acid change.

Table 2 | PolyBayes SNP and indel validation data

Mask type applied	Assay type	Submitted to validation	Assay successful	Sequencing successful	SNP candidate confirmed	Validation rate (%)	Conversion rate (%)
Known repeats	SNP	598	582	557	482	86.5	80.6
Exact microrepeats	SNP	579	559	518	475	91.7	82.0
Near-exact microrepeats (2 or fewer mismatches)	SNP	492	482	458	438	96.3	89.0
Known repeats	Indel	239	228	222	202	91.0	84.5
Exact microrepeats	Indel	232	223	217	201	92.6	86.6
Near-exact microrepeats (2 or fewer mismatches)	Indel	220	213	208	193	93.8	87.7

Validation and conversion rates for PolyBayes-selected SNPs and single base indel candidates. Successive application of masking filters, as described in the text, reduced the number of paralogous placements and identified high confidence putative variant sites.

DISCUSSION

Massively parallel sequencing approaches hold great promise for genome-wide discovery of sequence variation, when comparing different isolates or strains to reference genomes. It is apparent that short-read technologies must initially be characterized with respect to their quality and accuracy, providing a baseline for devising analytical methods. Dramatically shorter read lengths also increase the coverage level needed for adequate depth and breadth of reads to predict variation with high confidence, when compared to capillary sequencing reads. Although these short reads presently are too short for *de novo* assembly, producing regional assemblies of resequencing reads, followed by reference genome alignment, apparently has merit for detecting insertions and deletions, and should be pursued in future resequencing efforts.

Paired-end reads clearly increase the power to properly interpret problematic areas of the genome, including collapsed or misassembled repeats, and to detect structural variations. As genomes increase in size and complexity, paired ends will also be more efficiently placed than single-end reads, as only one end of each read pair needs a unique genome placement to properly place most reads, given that a precise paired-end read distance has been achieved in library construction.

Solexa reads provide a rapid vehicle for genome-wide SNP and small indel discovery, once additional masking of 'microrepeat' sequences is achieved. Aside from SNP or indel discovery, whole-genome resequencing also can be used after random mutagenesis to identify and characterize each mutagenized base. Our results establish the utility of short-read-length massively parallel sequencing for the accurate discovery of both single-nucleotide and small insertion-deletion polymorphisms, and establishes a framework for human genome resequencing toward similar discovery aims.

METHODS

Determining Solexa single-end read accuracy. To isolate sequencing errors from simple alignment errors, we used a version of the Smith-Waterman–based global alignment algorithm that reports all optimal and suboptimal alignments above a prespecified alignment score (EagleDiscover; W.H., unpublished data). Although time-intensive, this algorithm identifies all alignable positions in the *C. elegans* genome for a 32-bp read. Here we generated three random sample sets of 20,000 Solexa N2 Bristol single-end reads and aligned each read set to the unmasked reference genome, allowing up to 4 mismatches (substitution, insertion or deletion). For further consideration of accuracy, we kept only reads that aligned at a single locus in the genome. For each of the three read sets we tabulated the number of sequence differences between each read and the reference sequence, and combined the results to make a histogram of reads (**Fig. 2**). Then we evaluated unique alignments to calculate the observed error rate at each base position for a given Solexa base quality score. We converted these rates to phred scores (Solexa base qualities are expressed as a probability of each of the 4 bases being the correct call rather than as a single phred-like probability of correctness) and graphed the dependence of observed base quality on base position (**Fig. 3**).

Alignment and analysis of Solexa single-end reads. We compared Solexa N2 Bristol reads to the reference genome to identify sequence variants, to analyze coverage and to evaluate representational bias. These alignments consisted of a combination of exact hash-match based comparisons, followed by BLAST-like alignment tool (BLAT)-based comparisons. Our methods are detailed below, and are presented in a flowchart format (**Fig. 1** and **Supplementary Data**).

Paired-end read evaluation. We mapped paired-end reads (for example, a 25–35 bp read from each end of a ~200-bp genomic fragment) from the N2 Bristol isolate to the *C. elegans* genome using the exact hash-match based method described above. After read mapping of individual paired ends, we determined final placements by asking that the 'forward' and 'reverse' read of the pair match on the same chromosome and within 1,000 bases of each other.

Mosaik alignment of CB4858 Solexa reads. We identified both perfect microrepeats and microrepeats with up to two mismatches (substitutions, deletions or insertions) to encompass the possibility of sequencing errors (nucleotide misincorporation or base calling) in the reads or of polymorphism in the genomes being compared. Custom scripts then produced a microrepeat-masked reference genome.

We next aligned the Solexa CB4858 single-end reads to the microrepeat-masked *C. elegans* reference genome with our Mosaik program. Mosaik consists of two parts: the aligner (aligns each read to the reference genome separately in a pair-wise fashion) and the assembler (pads the individual reads and the reference genome sequence so that every aligned base within each read remains in register in the resulting multiple read alignment). The details of Mosaik processing are described in **Supplementary Methods**. The resulting multiple read alignments were then reported either in ACE[17] or in binary formats used by downstream analysis software.

SNP and indel discovery in strain CB4858. Starting with the multiple read alignments produced by the Mosaik aligner and assembler, we analyzed the resulting alignments using a version of PolyBayes[8] that was completely reengineered to enable efficient analyses of millions of aligned short-read sequences. The program evaluated each aligned base and its base quality value at each position, to indicate putative SNPs and small (1–3 bp) putative indels, and their corresponding SNP probability value (P_{SNP}). Base quality values were converted to base probabilities corresponding to every one of the four possible nucleotides (and to the probability that the nucleotide in question was an actual insertion error in the sequence). Using a Bayesian formulation[8], a P_{SNP} (or indel probability value, as appropriate) was calculated as the likelihood that multiple different alleles are present between the reference genome sequence and the reads aligned at that position. If the probability value exceeds a prespecified threshold, the SNP or indel candidate is reported in the output. For the collection of bases contributed by such reads, a single 'consensus' base call and its base quality value are computed. The corresponding base probabilities are then used in the Bayesian P_{SNP} calculation. In this study, we used a P_{SNP} cutoff value of 0.7 to define a high-certainty SNP or small indel site. Validated CB4858 SNPs and indels were assigned Wormbase accession numbers pas1– pas50906.

Software availability. The combined microrepeat plus RepeatMasker masked genome sequence annotations and FASTA files are available at http://bioinformatics.bc.edu/microrepeats/elegans/. Mosaik and the updated version of PolyBayes are now in beta release and available for users to wish to participate in software testing (http://bioinformatics.bc.edu/marthlab/Beta_Release). After the testing period, both programs will be released for public use, free of charge for academic users.

Additional methods. Details of Solexa library construction and sequencing, data analysis of primary sequence data and its alignment to the *C. elegans* reference genome (both single and paired-end reads) as well as detailed descriptions of Mosaik and PolyBayes analysis of CB4858 read data and its validation are available in **Supplementary Methods**.

Note: Supplementary information is available on the Nature Methods website.

ACKNOWLEDGMENTS

We acknowledge National Human Genome Research Institute funding (HG003079-04 to R.K.W. and HG003698 to G.T.M.). We thank K. Hall and D. Bentley of Illumina, Inc. for generously producing the paired-end read data described in the manuscript, M. Wendl for careful reading of the manuscript and T. Bieri for submitting the CB4858 variants to Wormbase.

AUTHOR CONTRIBUTIONS

L.W.H., N2 Bristol read, coverage, variant and gap analyses; G.T.M., CB4858 SNP discovery and N2 Bristol error profile analysis; A.R.Q., CB4858 SNP discovery and validation analysis; D.D., Solexa analysis pipeline; G.F., validation assay design and analysis, D.B., Solexa base quality value analysis, P.F., preparation of N2 Bristol and CB4858 DNA, J.I.G., N2 Bristol read analysis; M.H., Solexa libraries and sequencing, W.H., microrepeat analysis, V.J.M., Solexa libraries and sequencing, R.J.R., N2 Bristol analysis; S.N.S., validation assays; D.A.S., microrepeat masking of *C. elegans*; M.S., Mosaik adaptation; E.F.T., microrepeat finding; T.W., N2 Bristol analysis, T.S., *C. elegans* strain selection; R.K.W., project origination; E.R.M., project coordination and manuscript preparation.

Published online at http://www.nature.com/naturemethods/
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions

1. *C. elegans* Sequencing Consortium. Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science* **282**, 2012–2018 (1998).
2. Waterston, R. *et al.* The genome of the nematode *Caenorhabditis elegans*. *Cold Spring Harb. Symp. Quant. Biol.* **58**, 367–376 (1993).
3. Lander, E.S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
4. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature* **431**, 931–945 (2004).
5. Harris, T.W. *et al.* WormBase: a multi-species resource for nematode biology and genomics. *Nucleic Acids Res.* **32**, D411–D417 (2004).
6. Stein, L.D. *et al.* The genome sequence of Caenorhabditis briggsae: a platform for comparative genomics. *PLoS Biol.* **1**, e45 (2003).
7. Hodgkin, J. & Doniach, T. Natural variation and copulatory plug formation in *Caenorhabditis elegans*. *Genetics* **146**, 149–164 (1997).
8. Marth, G.T. *et al.* A general approach to single-nucleotide polymorphism discovery. *Nat. Genet.* **23**, 452–456 (1999).
9. Bieri, T. *et al.* WormBase: new content and better access. *Nucleic Acids Res.* **35**, D506–D510 (2007).
10. Tuzun, E. *et al.* Fine-scale structural variation of the human genome. *Nat. Genet.* **37**, 727–732 (2005).
11. Denver, D.R., Morris, K. & Thomas, W.K. Phylogenetics in *Caenorhabditis elegans*: an analysis of divergence and outcrossing. *Mol. Biol. Evol.* **20**, 393–400 (2003).
12. Smit, A.F. The origin of interspersed repeats in the human genome. *Curr. Opin. Genet. Dev.* **6**, 743–748 (1996).
13. Bhangale, T.R., Stephens, M. & Nickerson, D.A. Automating resequencing-based detection of insertion-deletion polymorphisms. *Nat. Genet.* **38**, 1457–1462 (2006).
14. Stephens, M., Sloan, J.S., Robertson, P.D., Scheet, P. & Nickerson, D.A. Automating sequence-based detection and genotyping of SNPs from diploid samples. *Nat. Genet.* **38**, 375–381 (2006).
15. Nickerson, D.A., Kolker, N., Taylor, S.L. & Rieder, M.J. Sequence-based detection of single nucleotide polymorphisms. *Methods Mol. Biol.* **175**, 29–35 (2001).
16. Koch, R., van Luenen, H.G., van der Horst, M., Thijssen, K.L. & Plasterk, R.H. Single nucleotide polymorphisms in wild isolates of *Caenorhabditis elegans*. *Genome Res.* **10**, 1690–1696 (2000).
17. Gordon, D., Abajian, C. & Green, P. Consed: a graphical tool for sequence finishing. *Genome Res.* **8**, 195–202 (1998).