Published in final edited form as:
Nat Methods. 2008 October ; 5(10): 865–867. doi:10.1038/nmeth.1249.

Caenorhabditis elegans mutant allele identification by whole-genome sequencing

Sumeet Sarin[1], **Snehit Prabhu**[2], **M Maggie O'Meara**[1], **Itsik Pe'er**[2], and **Oliver Hobert**[1]

1*Department of Biochemistry and Molecular Biophysics, Howard Hughes Medical Institute, Columbia University Medical Center, 701 W. 168th Street, New York, New York 10032, USA.*

2*Department of Computer Science, Columbia University, 1214 Amsterdeam Avenue, New York, New York 10027, USA.*

Abstract

Identification of the molecular lesion in *Caenorhabditis elegans* mutants isolated through forward genetic screens usually involves time-consuming genetic mapping. We used Illumina deep sequencing technology to sequence a complete, mutant *C. elegans* genome and thus pinpointed a single-nucleotide mutation in the genome that affects a neuronal cell fate decision. This constitutes a proof-of-principle for using whole-genome sequencing to analyze *C. elegans* mutants.

C. elegans is used extensively to identify genes involved in various aspects of animal development, behavior and physiology[1] (http://www.wormbook.org/). The traditional forward genetic approach involves random mutagenesis and subsequent isolation of mutants defective in a given process[1]. The ensuing characterization of the molecular lesion in a mutant strain is a painstaking process that involves mapping with genetic and/or single-nucleotide polymorphism (SNP) markers. The relative gene density in *C. elegans* and limited recombinant frequencies can make traditional mapping a very time-consuming process. This issue becomes even more apparent when the scoring of the mutant phenotype is cumbersome and recombinants are therefore tedious to identify. Another problem with traditional mapping approaches is that genetic-background effects on a given phenotype prohibit the use of many genetic markers and mapping strains.

We considered whole-genome sequencing as an approach to identify the molecular lesion in a specific ethyl methanesulfonate (EMS)-induced mutant *C. elegans* strain. We had previously described a genetic locus, *lsy-12*, in which a neuronal fate decision is aberrantly executed[2]. Instead of generating two left/right asymmetric, distinct chemosensory neurons, ASEL and ASER, *lsy-12* mutants generate two ASER neurons[2]. To determine the molecular identity of *lsy-12*, we undertook a single mapping cross of the recessive *lsy-12* allele *ot177* (*lsy-12* (*ot177*)) with a Hawaiian mapping strain[3], analyzed 200 F_2 progeny for SNPs by PCR and thereby mapped *lsy-12*(*ot177*) to a 4-Mb interval on chromosome V. This interval represents 4% of the genome and contains 1,142 predicted genes (5% of total genes).

We prepared genomic DNA from *lsy-12*(*ot177*) worms and sequenced the DNA using paired-end Illumina (formerly Solexa) sequencing technology[4]. We generated 4.35 Gb of paired 35-

Correspondence should be addressed to I.P. (ip2169@columbia.edu) or O.H. (or38@columbia.edu)..
AUTHOR CONTRIBUTIONS S.S. isolated and mapped *lsy-12*(*ot177*), performed the manual resequencing analysis and the RNAi analysis; S.P. performed the bioinformatic analysis; I.P. designed, performed and supervised the bioinformatic analysis; M.M.O. performed complementation tests and rescue analysis; and O.H. initiated and supervised the project and wrote the paper.

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

mer sequence reads in a 1-week sequencing run. Then we mapped the sequence data to the wild-type N2 reference genome using ELAND (efficient large-scale alignment of nucleotide databases) and the Maq alignment tools (Supplementary Methods online); 3.1 Gb of reads were mapped exactly onto the genome with an average coverage of ~28×. To label differences between our sequence data and the N2 reference genome as 'variants', we filtered for those reads that mapped uniquely with high-quality scores on both strands and were read at least ten times, thus eliminating the vast majority of ambiguous calls (Supplementary Methods). The filtering left 80 variants between *lsy-12*(*ot177*) genomic DNA and the published N2 wild-type reference genome in the 4-Mb interval, into which *lsy-12*(*ot177*) mapped. We ranked these 80 variants according to standard quality scores (Supplementary Table 1 online). Fifty-four of the 80 variants were single-nucleotide variants and 26 were small, mainly 1-nt insertions-deletions (indels; Fig. 1 and Supplementary Table 1). None of the indels mapped to exons or splice sites of predicted genes, and 21 of the 54 single-nucleotide variants affected exons of protein-coding genes. Five of the 21 exonic variants were silent variants. The remaining 16 variants (15 missense and 1 nonsense) were the best candidates for the *lsy-12* mutation as more than 90% of EMS-induced mutant alleles are generally point mutations that introduce changes in amino acids or splice junctions (Supplementary Table 2 online).

To determine whether the detected variations were (i) sequencing or mapping errors in the Illumina Genome Analyzer pipeline, (ii) sequence variations in the original transgenic strain mutagenized, a strain containing the transgene *otIs114* (ref. 5) or (iii) true mutagenesis-induced mutations, we PCR-amplified ~300-bp fragments that contained each of the 80 variants from both the starting strain used for mutagenesis (strain containing the transgene *otIs114)*, and from the mutant strain (*lsy-12*(*ot177*)), which also contains the unlinked *otIs114* transgene. Each of these strains had been outcrossed against N2 wild-type multiple times. Resequencing of the amplicons derived from the *lsy-12*(*ot177*) mutant strain by the traditional Sanger method confirmed the presence of all of the 26 indels. Notably, all of the indels were already present in the transgenic starting strain containing the transgene *otIs114*. We then Sanger-sequenced the genomic regions containing these indels in our available N2 wild-type isolate and found each of the indels to be present in N2 as well. We therefore unintentionally uncovered a large amount of sequence variation between the N2 reference genome and the N2 wild-type strain distributed by the Caenorhabditis Genetics Center at the University of Minnesota, which could be due to genetic drift or to errors in sequencing the reference genome. These findings are consistent with a previous study[6].

We confirmed 17 of the 33 non-exonic variants by Sanger-sequencing of *lsy-12*(*ot177*) (Fig. 1). The remaining 16 uncon-firmed 'variants' (10 were apparent sequencing errors and 6 were clustered within a single intron that we could not reliably identify in the genome) were supported by less variant reads than wild-type reads, thereby suggesting another rule by which to filter variants (Supplementary Table 1a). Of the 17 confirmed variants, 8 variants were present in both the starting strain containing the transgene *otIs114* and also in our N2 wild-type strain, again underscoring the sequence differences between the reference genome and our wild-type strain (Fig. 1 and Table 1).

We confirmed 15 out of the 21 exonic variants, and specifically 11 of the 16 exonic variants that altered an amino acid, by Sanger resequencing (Supplementary Table 1a). The remaining erroneous variants again had lower quality scores and in each case, we observed more wild-type than variant reads (Fig. 1 and Supplementary Table 1a). Of the 11 confirmed exonic variants that alter an amino acid, 7 were already present in the starting strain containing the trans-gene *otIs114* and most of these again were present in our N2 wild-type isolate (Fig. 1). This left 4 amino acid—changing variants between the *lsy-12*(*ot177*) mutant and the N2 wild-type genome in the mapped 4-Mb interval (Fig. 2). In sum, we discounted the majority of initial variations in the mapped 4-Mb interval (80 variants) as they did not affect protein-coding

regions (64 of original 80 variants, 80%) and/or were sequencing errors and/or are variations between strain backgrounds, leaving only a total of four exonic variants that we predicted to alter a protein product (Fig. 1 and Table 1).

One of these four exonic variants is a nonsense mutation in the predicted *R07B5.9* gene, the sole nonsense mutation in the entire dataset. We sequenced this predicted gene in the other five available strains that harbor mutant alleles of *lsy-12*, as determined by complementation testing and mapping (Supplementary Methods and Supplementary Table 3 online). We found that each one of them harbors a mutation in *R07B5.9* (Fig. 2). None of the *lsy-12* strains displayed variations in the other three candidate genes revealed by genome sequencing. Moreover, the *lsy-12* mutant phenotype was rescued by injecting a ~39 kb genomic interval that contains *R07B5.9* but no other candidate gene suggested by the whole-genome analysis (Fig. 2 and Supplementary Table 3). Lastly, we performed RNA interference (RNAi) of all four genes with exonic variants and found that only RNAi of *R07B5.9* phenocopies the *lsy-12* mutant phenotype (Supplementary Table 3). We concluded that *lsy-12* is *R07B5.9*. Whole-genome sequencing has therefore revealed the identity of a previously unknown mutant gene.

The ability to perform additional experiments to distinguish the true phenotype-causing mutation from the set of sequence variants identified by whole-genome sequencing will dictate the amount of mapping one needs to perform before using whole-genome sequencing. From our identification of four protein-changing variants in a 4-Mb interval, we extrapolate that an entire chromosome, such as chromosome V (20.9 Mb), may only contain ~20 protein-changing candidate variants. The availability of multiple alleles is the easiest way to sift through these candidates, as it is fast and simple to manually Sanger-sequence many candidate genes in the allelic strains. RNAi and transformation rescue represent other powerful tools to test whether sequence variants are responsible for the mutant pheno-type. We conclude that minimal mapping to as little as a chromosome and perhaps even less is required before using a whole-genome sequencing strategy.

To facilitate the design of future studies, we statistically analyzed the sequence data and found that the number of reads at a particular location (that is, the coverage), did not follow traditional formulae[7] but could be approximated as a gamma-distributed random variable (see Supplementary Methods and Supplementary Table 4 online). Under the assumption of our observed 0.6% error rate, this observation predicts that 8× coverage would yield ~150 variants in a 4-Mb interval, supported by at least four variant reads (see Supplementary Methods for details on these predictions and their detection power). Given that, in our case, only 4 in 80 variants in a 4-Mb interval were non-silent variants within a coding region (Fig. 1), this would translate into only ~8 variants requiring validation. We therefore recommend aiming for eightfold coverage, which should be reached with ~0.8 Gb of aligned sequence, produced by two lanes in a Genome Analyzer flow cell.

The application of whole-genome sequencing offers many unique advantages. The sequence run only takes a few days and costs are in the few-thousand dollar range, which compare favorably to the personnel and reagent costs of a traditional multi-year gene cloning project. The implications of the substantial time savings with this approach go beyond mere cost considerations. The approach should motivate large-scale genetic screens, followed by the rapid sequencing of many mutants retrieved from such screens. This will not only lead to a more comprehensive genetic understanding of a given biological process but will offer the practical advantage of being able to sift through a collection of mutants and to focus on those genes whose molecular identity bear the most interest to the investigator. Moreover, mutants for which the phenotype is tedious to score (for example, behavioral mutants), mutants for which the phenotype is subject to modification by the genetic background of mapping strains

and mutants that require specific genetic backgrounds (that is, modifier mutants), can now be more easily identified by whole-genome sequencing.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGMENTS

We thank C.T. Lawley (Illumina, Inc.) for generously producing the paired-end read data described in the manuscript, Q. Chen for performing microinjection, Hobert lab members for comments on the manuscript and the Caenorhabditis Genetics Center for providing the N2 strain. This work was funded by the Howard Hughes Medical Institute, the Muscular Dystrophy Association and the US National Institutes of Health (R01NS039996-05; R01NS050266-03 to O.H., F31 predoctoral grant NS054540-01 to S.S. and U54 CA121852-03 to I.P.).

References

1. Brenner S. Genetics 1974;77:71–94. [PubMed: 4366476]
2. Sarin S, et al. Genetics 2007;176:2109–2130. [PubMed: 17717195]
3. Davis MW, et al. BMC Genomics 2005;6:118. [PubMed: 16156901]
4. Bentley DR. Curr. Opin. Genet. Dev 2006;16:545–552. [PubMed: 17055251]
5. Chang S, Johnston RJ Jr. Hobert O. Genes Dev 2003;17:2123–2137. [PubMed: 12952888]
6. Hillier LW, et al. Nat. Methods 2008;5:183–188. [PubMed: 18204455]
7. Lander ES, Waterman MS. Genomics 1988;2:231–1239. [PubMed: 3294162]

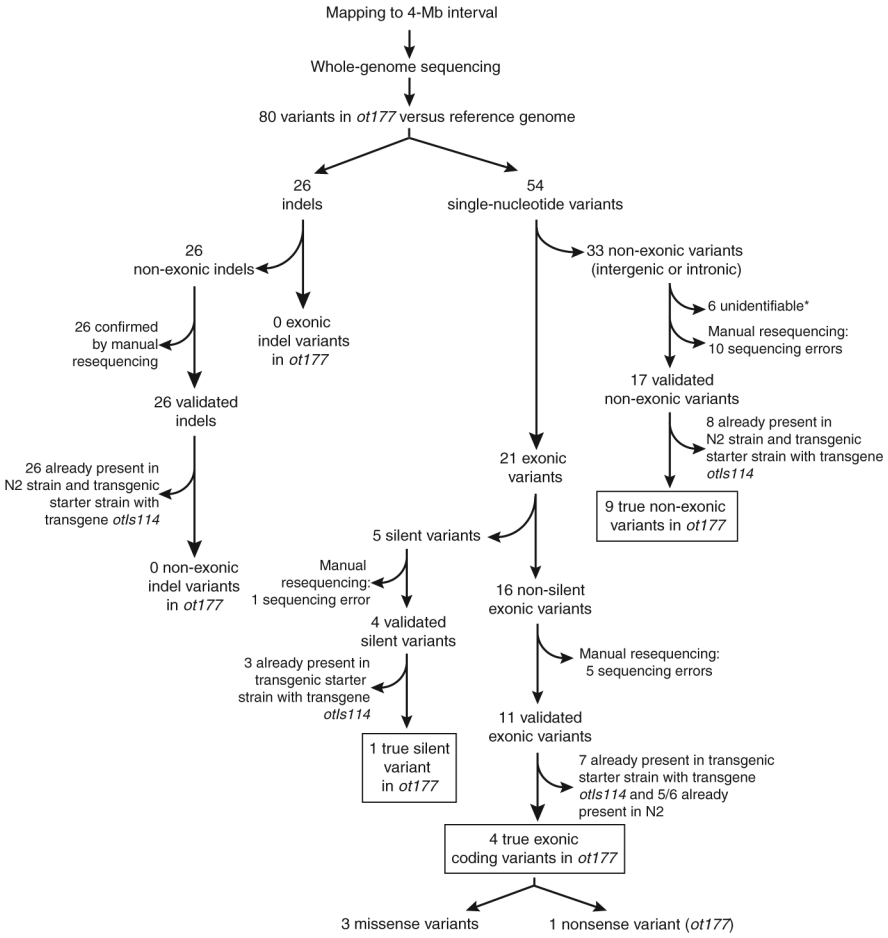

Figure 1.

Variants found in whole-genome sequence analysis. *ot177* is shorthand for *lsy-12*(*ot177*). The asterisk denotes 6 variants clustered within 100 bp of a single intron (supplementary Table 1a), which upon amplification and Sanger-sequencing we found to map at least in part onto a different chromosome.

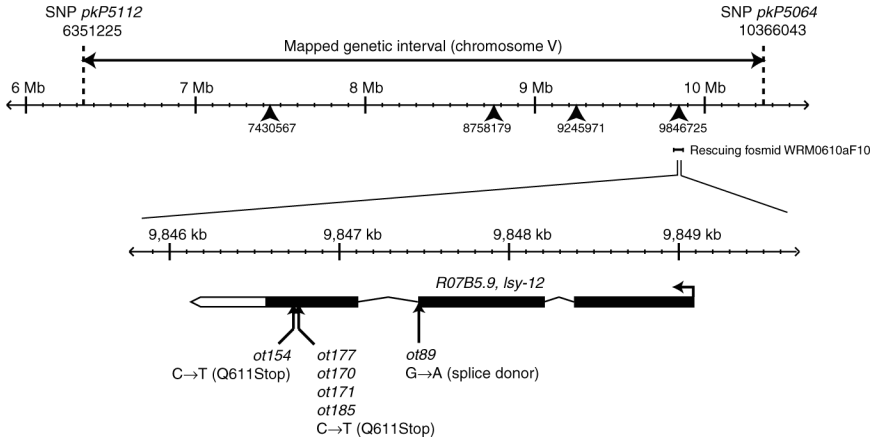

Figure 2.

Physical location of the *lsy-12* locus and its lesions. *lsy-12*(*ot177*) was mapped between two SNP markers, pkP5112 and pkP5064 (http://www.wormbase.org/), which define a 4-Mb interval. Validated exonic variants in this interval are marked with arrowheads. Numbers indicate base position on chromosome V. The gene *R07B5.9*, identified as *lsy-12*, is shown below and the positions of the previously identified *lsy-12* alleles are marked with black arrows.

Table 1

Frequency of sequence variation types

Source of variation	Frequency
Sequencing errors in Genome Analyzer-identified variations over total variations	21.6% (10 + 5 + 1) / (80 − 6[a])
True variations between starting strain (strain containing the transgene *otIs114*) and mutant strain (*lsy-12*(*ot177*); *otIs114*) over total variations	18.9% (9 + 1 + 4) / (80 − 6[a])
Variations due to differences between available N2 strain and reference genome over total variations	56.8% (26 + 8 + 3 + 5) / (80 − 6[a])

[a]Six variants clustered within 100 bp of a single intron (Supplementary Table 1a), which upon amplification and Sanger sequencing we found to map at least in part onto a different chromosome. We subtracted these six variants in the summary calculations as their source was not clear.