Allelic Ratios and the Mutational Landscape Reveal Biologically Significant Heterozygous SNVs

Jeffrey S.-C. Chu,*,†,1 Robert C. Johnsen,† Shu Yi Chua,† Domena Tu,*,† Mark Dennison,† Marco Marra,‡ Steven J. M. Jones,*,†,‡ David L. Baillie,† and Ann M. Rose*

*Department of Medical Genetics, University of British Columbia, Vancouver, BC V6T 1Z4, Canada, †Molecular Biology and Biochemistry, Simon Fraser University, Burnaby, BC V5A 1S6, Canada, and ‡Genome Sciences Centre, British Columbia Cancer Research Centre, Vancouver, BC V5Z 4E6, Canada

ABSTRACT The issue of heterozygosity continues to be a challenge in the analysis of genome sequences. In this article, we describe the use of allele ratios to distinguish biologically significant single-nucleotide variants from background noise. An application of this approach is the identification of lethal mutations in *Caenorhabditis elegans* essential genes, which must be maintained by the presence of a wild-type allele on a balancer. The *h448* allele of *let-504* is rescued by the duplication balancer *sDp2*. We readily identified the extent of the duplication when the percentage of read support for the lesion was between 70 and 80%. Examination of the EMS-induced changes throughout the genome revealed that these mutations exist in contiguous blocks. During early embryonic division in self-fertilizing *C. elegans*, alkylated guanines pair with thymines. As a result, EMS-induced changes become fixed as either G→A or C→T changes along the length of the chromosome. Thus, examination of the distribution of EMS-induced changes revealed the mutational and recombinational history of the chromosome, even generations later. We identified the mutational change responsible for the *h448* mutation and sequenced PCR products for an additional four alleles, correlating *let-504* with the DNA-coding region for an ortholog of a NFκB-activating protein, NKAP. Our results confirm that whole-genome sequencing is an efficient and inexpensive way of identifying nucleotide alterations responsible for lethal phenotypes and can be applied on a large scale to identify the molecular basis of essential genes.

FORWARD genetics in model organisms, which involves random mutation and isolation of a phenotype, laid the foundation for characterization of gene function. The bottleneck of this process lies in the identification of the molecular lesion responsible for the phenotype. The traditional approach for mutation identification involves three-factor mapping followed by several rounds of complementation testing using deficiencies and duplications. To reduce the number of candidate-coding regions, cosmids and fosmids are used to attempt to rescue the lethal phenotype (Janke *et al.* 1997; Simms and Baillie 2010). Finally, PCR analysis and DNA sequencing are used to confirm the molecular identity of the gene. This approach is laborious, time-consuming, and has very low throughput.

doi: 10.1534/genetics.111.137208
Manuscript received November 27, 2011; accepted for publication January 10, 2012
Available freely online through the author-supported open access option.
Supporting information is available online at http://www.genetics.org/content/suppl/2012/01/20/genetics.111.137208.DC1.
[1]Corresponding author: University of British Columbia, Room 419, NCE Bldg., 2125 East Mall, Vancouver, BC V6T 1Z4, Canada. E-mail: jeff.sc.chu@gmail.com

Technological advancements have provided methods to speed up the process of mutation identification. Recently, array comparative genomic hybridization (aCGH) was applied to identify single-nucleotide variations (SNVs) in the genomes of *Saccharomyces cerevisiae* (Gresham *et al.* 2006) and *Carnorhabditis elegans* (Maydan *et al.* 2009). This genome-wide approach allows rapid identification of a region of interest without mapping the mutation. Together with dense tiling arrays, aCGH could narrow down a SNV to within 10 bp (Maydan *et al.* 2009). However, this approach, which relies on sensitive hybridization, is unable to detect heterozygous mutations (Gresham *et al.* 2006; Maydan *et al.* 2009).

Whole-genome sequencing (WGS) is coming to the forefront as an attractive alternative for identifying molecular lesions (Cronn *et al.* 2008; Hobert 2010). Many researchers, including ourselves, have successfully identified SNVs and large genomic variations using WGS (Sarin *et al.* 2008, 2010; Shen *et al.* 2008; Doitsidou *et al.* 2010; Flibotte *et al.* 2010; Maydan *et al.* 2010; Rose *et al.* 2010). This approach has greatly facilitated the characterization of mutant phenotypes as well as many natural variants (Hillier *et al.* 2008). WGS is particularly

useful for identifying hard-to-map alleles and genes that cannot be rescued by conventional transgenic fosmid or cosmid libraries. Nevertheless, almost all of the studies to date have focused on identifying homozygous mutations whereas identifying heterozygous mutations continues to be a challenge. Identifying heterozygous SNVs is an important step in genome analysis for understanding genomic variations and is generally relevant to many situations where allelic differences exist. In this report, we have developed a method for effectively identifying heterozygous mutations in *C. elegans* using lethal mutations as a model.

C. elegans is a self-fertilizing hermaphrodite whose genome becomes homozygous within a few generations. However, *C. elegans* with lethal mutations cannot be maintained as viable homozygous strains. This problem has been solved by the use of "balancers" to isolate, maintain, and characterize essential gene mutations (Edgley *et al.* 2006). In *C. elegans*, two commonly used classes of balancer are translocations (Rosenbluth and Baillie 1981) and duplications (Rose *et al.* 1984). Duplications that do not crossover with the normal chromosomes provide a third allele that is wild type and can rescue the lethal mutation, which is effectively maintained as a homozygote. In the case of the duplication-rescued strains, the allele frequency is 2:1 mutant:wild type. In this article, we describe the first use of Illumina sequencing to identify the DNA-coding region of an essential gene rescued by the duplication *sDp2* on chromosome I of *C. elegans*.

Materials and Methods

C. elegans strains

C. elegans is a self-fertilizing hermaphrodite and produces isogenic progeny within a few generations. Strains carrying homozygous mutations in *let-504* rescued by the duplication *sDp2* were previously generated (Howell *et al.* 1987) and complementation tested (Howell and Rose 1990). The wild-type N2 (KR4848) is a derivative of a CGC N2 strain maintained in the Rose laboratory. The strain carrying a heterozygous *tm4719* allele was kindly provided by S. Mitani (National Institute of Genetics, Mishima, Shizuoka, Japan). The strain KR5173, which carries the *tm4719* allele balanced by *hT2[bli-4(e937)] let-x (q782) qIs48]* I; III, was generated in this study. *C. elegans* strains were maintained at 20° as previously described (Brenner 1974).

Genomic DNA preparation

KR772 and KR4848 worms were grown on five 10-cm agar plates with *Escherichia coli* OP-50 until food was depleted (~5 days at 20°). Worms were collected and pelleted by washing the plate with M9 and centrifuged at 1500 × *g* for 1 min at 4°. The worm pellet was washed three times with M9, followed by 2–3 hr of incubation at room temperature to allow bacteria digestion. The worms were pelleted as before and finally resuspended in 0.5 ml of TE. The worms were frozen in −20° and lysed in lysis solution (50 μl 5% SDS, 2.5 μl 20 mg/ml Proteinase K) at 60° for 2 hr. Genomic DNA was purified using phenol/chloroform extraction and ethanol precipitation. The sample was treated with 4 μl of 5 mg/ml RNase A for 1 hr at 37°, followed by a second round of phenol/chloroform extraction and ethanol precipitation. Approximately 10 μg of DNA was sheared for 10 min using Sonic Dismembrator 550 (cup horn, Fisher Scientific) with a power setting of "7" for 30-sec pulses interspersed with 30 sec of cooling and analyzed on a 8% PAGE gel. A 180- to 220-bp DNA fraction was excised and eluted from the gel slice overnight at 4° in 300 μl of elution buffer [5:1, LoTE buffer (3 mM Tris–HCl, pH 7.5, 0.2 mM EDTA):7.5 M ammonium acetate] and was purified using a Spin-X Filter Tube (Fisher Scientific) and by ethanol precipitation. The whole genome shotgun sequencing library was prepared using a modified paired-end protocol supplied by Illumina. This involved DNA end-repair, formation of 3′ A overhangs using Klenow fragment (3′–5′ exo minus), and ligation to Illumina PE adapters. Adapter-ligated products were purified on Qiaquick spin columns (Qiagen) and PCR-amplified using Phusion DNA polymerase for 10 cycles using the PE primer 1.0 and 2.0 (Illumina). PCR products of the desired size range were purified using a 8% PAGE gel. DNA quality and quantity was assessed using an Agilent DNA 1000 series II assay and Nanodrop 7500 spectrophotometer (Nanodrop), and DNA was subsequently diluted to 10 nM. The final concentration was confirmed using a Quant-iT dsDNA HS assay kit and Qubit fluorometer (Invitrogen). For sequencing, clusters were generated on the Illumina cluster station and paired end-reads were generated using an Illumina GAII platform following the manufacturer's instructions. Image analysis, base calling, and error calibration was performed using the V1.0 Illumina Genome Analyzer analysis pipeline.

Mutational density calculation

We collected coordinates of all the homozygous EMS changes. The genome was divided into overlapping bins of 2 Mbp, and we counted the number of EMS changes in each bin. The mutational rate for each bin was derived by dividing the number of EMS changes by the bin size. The value for each bin was collected and used to plot Figure 4.

Whole-genome sequencing and analysis

The genomic sequence of KR772 was aligned to the annotated sequence of *C. elegans* available at WormBase WS200 (http://www.wormbase.org) using BWA at the default setting (Li and Durbin 2009) and compared with the sequence of the wild-type strain KR4848. Genome analysis and visualization were done using Integrative Genomics Viewer (Robinson *et al.* 2011). The SNVs were called using VarScan (Koboldt *et al.* 2009) with the following parameters: –min-coverage 20–min-avg-qual 20–min-var-frequation 0.2–p-value 0.1. Candidate nucleotide differences for *let-504* (*h448*) were further filtered to satisfy the following three criteria: (1) mutations that fall within genetic mapping range; (2) unique to mutant strain compared to the N2 strain (KR4848); and (3) allelic ratio falls between 60 and 90%.

Fosmid transgenic rescue

The fosmid WRM0614bH01 was injected into *dpy-5* hermaphrodites to construct the transgenic strain BC8626, which carries *myo-2*::GFP physically linked to *dpy-5(+)* and the fosmid. GFP males from BC8626 were mated to KR772, and individual GFP outcross hermaphrodites were isolated. The progeny of individual GFP hermaphrodites were examined for fertile GFP Unc animals, which could be (1) crossovers between the *let-504* and *dpy-5*, (2) carrying both the fosmid and *sDp2*, or (3) fosmid rescues of *let-504*. In the third case, GFP Unc animals would continue to segregate Uncs (all with GFP) and arrested Dpy Uncs. The length of the GFP Uncs was measured, and gonadal indexing was used to test for a shift in the *h448* phenotype.

Complementation test

We received a knockout allele of E01A2.4, *tm4719*, from S. Mitani (National Institute of Genetics). Animals homozygous for *tm4719* are sterile as adults. From a mixed population, animals heterozygous for the *tm4719* deletion allele were crossed with males heterozygous for *hT2[bli-4(e937)] let-x (q782) qIs48]* I; III, which carries an insertion of *myo-2*::GFP. GFP hermaphrodites were selected and tested for *tm4719* by PCR. One hermaphrodite carrying *tm4719* was used to establish a balanced strain. GFP *tm4719* males from the above cross were individually crossed to a single KR772 hermaphrodite [*sDp2*; *let-504(h448) dpy-5(e61) unc-13 (e450)/let-504(h448) dpy-5(e61) unc-13(e450)*]. The F_1 progeny from these individual crosses were screened for sterile non-Dpy non-Unc non-GFP adults. If *h448* fails to complement *tm4719*, we would expect sterile adults segregating in the outcross progeny. The *tm4719* deletion in sterile non-Dpy non-Unc non-GFP adults was tested with the PCR primers.

Results and Discussion

Mutant strain selection

Identification of the molecular basis of lethal mutations is problematic for WGS because the animals cannot be grown as homozygotes in large amounts for DNA production. In selecting a mutant strain to characterize, we took into account how well-mapped the mutation was and the number of alleles that failed to complement it. We chose the *h448* mutation that is in the essential gene *let-504*. The *h448* allele is maintained as a homozygote by a rescuing wild-type allele on the duplication *sDp2*. The free (unattached to a normal chromosome) duplication, *sDp2*, covers ~7.3 Mbp of the left half of chromosome I (Howell *et al.* 1987). We chose a deleted interval, *hDf7*, that is in the *sDp2*-balanced region because it was mapped to a well-defined area of ~200 kbp and contained a small number of essential genes (Figure 1). Of the six complementation groups mapping within *hDf7* (*let-353*, *let-503*, *let-504*, *let-505*, *let-506*, and *let-507*), *let-504* had the most alleles (Table 1) (Johnsen *et al.* 2000). It was with these considerations in mind that the strain KR772, which carries *let-504 (h448)*, was selected. Previous analysis showed that the phenotypes of the *let-504* alleles ranged from larval arrest to sterile adults (Howell *et al.* 1987; Johnsen *et al.* 2000). Our strategy was to identify the *let-504*-coding region by inspection of the genome sequence in the *hDf7* region and to validate its identity using DNA sequencing of PCR products from the additional alleles.

Nonrandom distribution of G→A and C→T changes

Genomic DNA of KR772 was prepared and sequenced using Illumina sequencing. For comparison, we prepared and sequenced the genome of KR4828, a Bristol wild-type (N2) strain from the Rose laboratory. In KR772, a total of 45,694,133 read pairs of 114 bp read length were generated. Approximately 87% of the reads were aligned to the annotated *C. elegans* genome (WS200) using BWA (Li and Durbin 2009). The number of reads provided ~80-fold coverage on average.

To better identify candidate mutations for *let-504*, we first analyzed the general mutational load of KR772. We compiled all the base-pair differences unique to KR772 using Varscan (Koboldt *et al.* 2009) (Supporting Information, Table S1). We observed 648 SNVs present with >90% read support and, of these, 55% (357) were either G→A or C→T changes (Figure S1), which are characteristic of EMS mutations (Bautz and Freese 1960; Greene *et al.* 2003). Even though the lethal mutation was induced using a relatively low dose of 15 mM EMS (Howell *et al.* 1987), compared to the 50-mM dose that is often used (Brenner 1974; Sulston and Hodgkin 1988; Flibotte *et al.* 2010; Sarin *et al.* 2010), there still appear to be a large number (357) of apparent EMS-induced changes across the whole genome.

We analyzed the positions of the homozygous SNVs in KR772 and found that these changes do not distribute evenly across the genome. Figure 2 shows the positions of G→A and C→T changes plotted separately along the chromosome. Surprisingly, EMS-induced mutations clustered in contiguous blocks of either G→A or C→T changes. In some cases, the blocks spanned the entire chromosome. We observed that on chromosomes II, V, and X, the changes are predominantly G→A, whereas those on chromosome I are predominantly C→T (Figure 2).

We explain our observations in the following way. The affected gametes of the EMS-treated parent will have some alkylated G's. For simplicity, we consider only the alkylated G's in the sperm (Figure 3A). In the first round of replication of embryonic cell division after fertilization, the alkylated G's will mis-pair with T's (Figure 3B). In the second round of replication, the T's will pair with A's (Figure 3C). This results in EMS mutations becoming fixed such that the alkylated G's have been replaced with A's. The mutational changes will be the same for the entire DNA strand. For example, the alkylated G's from the plus strand in the gamete will be fixed as A's, and the C's (G's from the minus strand) will be fixed as T's (Figure 3C). Only one of these

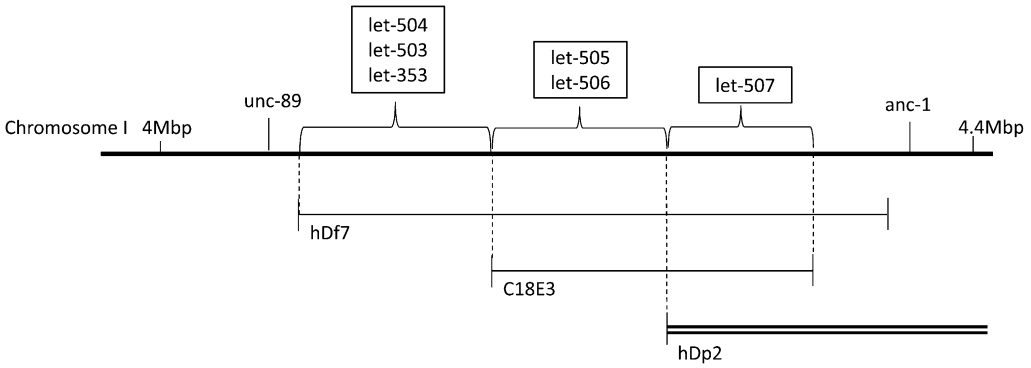


Figure 1 A map of lethal genes on chromosome I exposed by the deletion *hDf7*, which is in the *sDp2* region. Six lethal genes fall within this region. Three of these genes (*let-353*, *let-503*, and *let-504*) fall in the region flanked by the left breakpoint of *hDf7* and the left breakpoint of the cosmid C18E3.

possibilities will be segregated to the germ-line progenitor cell (P-cell lineage) and passed onto the next generation. Thus, the offspring will inherit either G→A changes or C→T changes along the entire chromosome for any one affected gamete. To test the generality of this observation, we examined available EMS-treated genomic sequences published by Flibotte *et al.* (2010). In the strains RB5002, VC1923, and VC1924, the G→A and C→T changes occur in long contiguous blocks similar to our observation (Figure S2).

We also observed both in our data and in data from Flibotte *et al.* (2010) that there is a shift from one block of EMS type to another within a chromosome (Figure 2 and Figure S2). For example, chromosome IV in KR772 has a stretch of G→A changes and shifts to a stretch of C→T changes. Similarly, we observed in chromosome V of VC1924 (as an example) where the EMS mutations shift from a block of C→T changes to a block of G→A changes and then back to C→T changes. The shift between blocks can be explained as a result of crossing over during meiosis and subsequent homozygosis in the self-progeny of the hermaphrodite. In these cases, the paternal chromosome may have contained only G→A changes and the maternal chromosome only C→T changes. Crossing over between the homologs during meiosis would result in a chromosome with a segment of G→A changes and another segment of C→T changes. In summary, WGS of EMS-treated strains provides a way of identifying the type of mutational change along large stretches of the chromosome.

In addition to differences in SNV types, we also observed differences in SNV density. A sparsely mutated chromosome would have a flatter slope whereas a densely mutated chromosome would have a steeper slope. We observed many chromosomes shift from a densely mutated segment to a sparsely mutated segment, or *vice versa* (Figure 2 and Figure S2). We calculated the density of EMS mutations (see *Materials and Methods*) and observed that a lower density of mutation averaged about two SNVs per mega base pair and a higher density of mutation averaged between four and six SNVs per mega base pair (Figure 4). The shift in mutational density along the chromosome is likely a result of meiotic crossing over between the paternal and maternal chromosome. If so, the mutational frequency in sperm differs from the frequency in oocytes. There is evidence for this difference in other species. In *Drosophila*, EMS is more effective when fed to males than to females (Lewis and Bacher 1968), suggesting that sperms have a higher mutational frequency. More effective repair mechanisms and an increased cytosolic volume in the oocyte that may act as a sink for alkylating agents could result in a lower mutation frequency. Thus, it is possible that hermaphrodite sperm are more sensitive than oocytes to mutation.

Identification of heterozygosity using allelic ratio

The left half of chromosome I has notably fewer homozygous mutations than the rest of the chromosome. We predicted that the SNVs in that region would have <90% read support due to heterozygosity. We counted the number of SNVs as a function of their allelic ratio (Figure 5). For a typical chromosome, most of the SNVs fall within a 90–100% allelic ratio (*e.g.*, the green line in Figure 5). However, in chromosome I, we observed a bi-modal distribution of the SNVs, with one peak at 70–80% and another peak at 90–100% (black line in Figure 5). Nearly all of the SNVs in the 70–80% category are located in the *sDp2* region (blue dashed line in Figure 5) whereas SNVs outside of the *sDp2* region are within 90–100% (red dashed line in Figure 5). We conclude that the EMS-induced mutations are homozygous along chromosome I homologs and differ from the wild-type alleles on the duplication, resulting in an allelic ratio in the range of 70–80%.

Table 1 Alleles of *let-504*

Strain	Allele	Mutagen	Genotype	Arrest stage
KR456	*h137*	EMS	*sDp2; let-504 (h137) dpy-5 (e61) unc-13 (e450)*	Sterile adult
KR661	*h327*	Gamma radiation	*sDp2; let-504 (h327) dpy-5 (e61) unc-13 (e450)*	L2/L3
KR772	*h448*	EMS	*sDp2; let-504 (h448) dpy-5 (e61) unc-13 (e450)*	Sterile adult
KR1506	*h844*	EMS	*sDp2; let-504 (h844) dpy-5 (e61) unc-13 (e450)*	L3
KR1541	*h888*	EMS	*hT1; let-504 (h888) dpy-5 (e61) unc-13 (e450)*	L2/L3

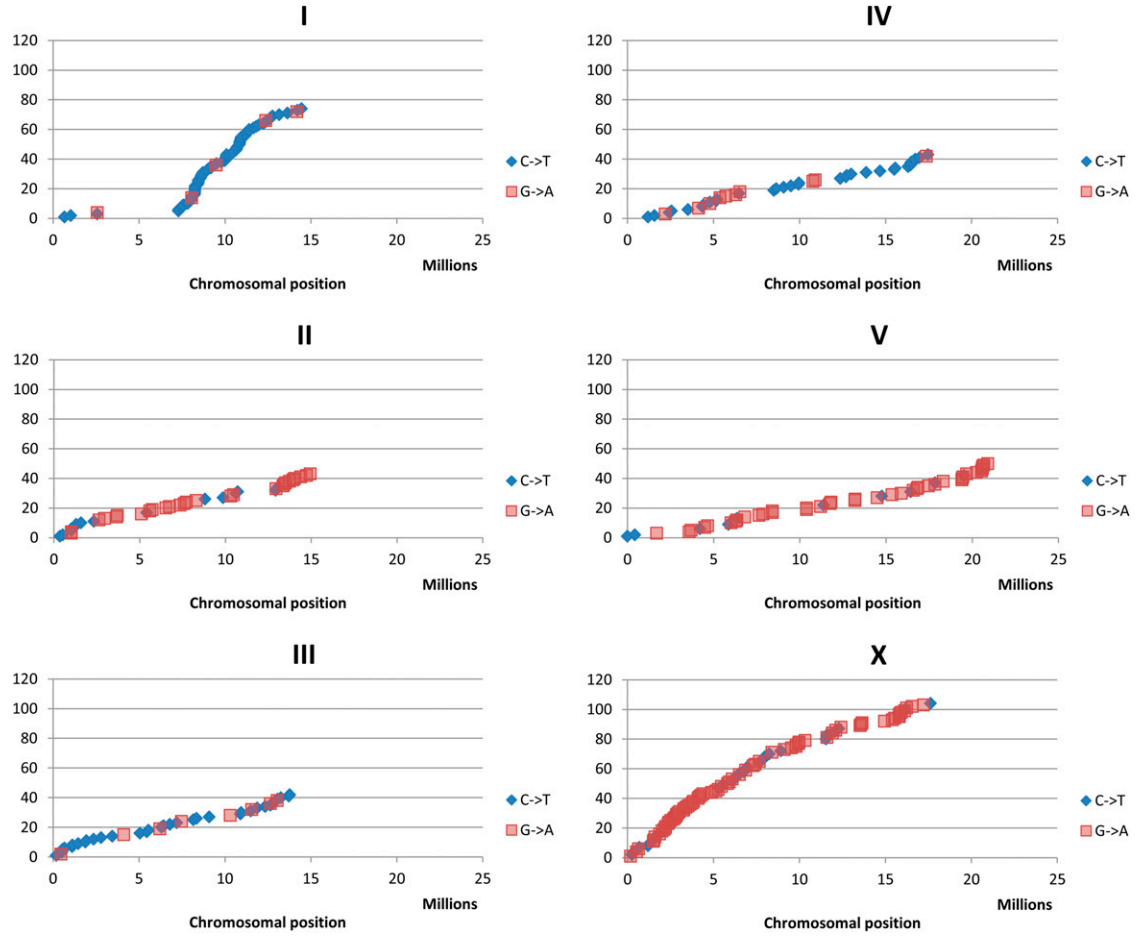


Figure 2 Positions of G→A (red squares) and C→T (blue diamonds) homozygous SNVs. Any SNV with >90% read support is considered homozygous. The *x*-axis represents the length of the chromosome in mega base pairs. The *y*-axis indicates each SNV ID. The scale is the same for all the chromosomes so that the slope of the line corresponds to the density of SNVs. Chromosomes I and III are predominantly C→T changes whereas chromosomes II, V, and X are predominantly G→A changes. Chromosome IV has a stretch of G→A changes followed by a stretch of C→T changes. Chromosome I and X have a steeper slope, indicating a higher density of EMS mutations.

We reasoned that we could use the allelic ratio to determine the extent of the duplication. We plotted the distribution of allelic ratios in 1-Mbp intervals along the chromosomes (Figure S3). Chromosomes I-right, II, III, IV, V, and the X are predominantly homozygous for the SNVs whereas the left half of chromosome I is predominantly non-homozygous for the SNVs (Figure S3). On chromosome I, homozygous SNVs emerge between 7 and 8 Mbp, indicating that the *sDp2* boundary falls within this interval. A detailed examination of the region (Figure 6) shows that homozygous mutations emerge between 7.2 and 7.3 Mbp, corresponding to the location of the right breakpoint of *sDp2*.

Application of EMS mutation blocks and allelic ratios to identify candidate mutations in the hDf7 region

The *hDf7* region contains 62 predicted protein-coding sequences. Of these predicted coding sequences, 43 are potentially essential genes by the observation of arrest phenotypes after RNA interference treatment (Fraser *et al.* 2000; Sonnichsen *et al.* 2005). Even with this shortened candidate list, identifying the molecular lesion for *let-504* among the candidates is a daunting task. WGS of a strain carrying one of the alleles therefore presents a viable approach to identifying the responsible mutation. To narrow down the candidates, we took advantage of the two genomic features discussed above: (1) EMS mutation types occur in contiguous blocks and (2) the allelic ratio is likely to be close to 70–80%. Our analysis of the EMS mutations showed that chromosome I is predominantly C→T changes, and thus we predicted that *let-504 (h448)* is also likely to be a C→T change.

The strain KR772 carries flanking markers in addition to the lethal mutation *let-504 (h448)*, and we examined the sequence for these pre-existing mutations. One of the markers, *dpy-5 (e61)*, is situated in the *sDp2* region, and the duplication provides a wild-type copy of the *e61* mutant allele. At position 5,432,448 on chromosome I, 80% (45/56 reads) had an A whereas the remaining 20% of the reads had a C at this position, which is the nucleotide in the wild-type N2 sequence. Our results are in agreement with the

A Alkylated G's shown in red

GAACAGTTTCTGTTC Plus strand in sperm
CTTGTCAAAGACAAG Minus strand

B At first replication alkylated G pairs with T

GAACAGTTTCTGTTC Plus strand from sperm
TTTGTTAAAGACAAG New Minus (note new T's)

GAACAGTTTTTGTTT New Plus(note new T's)
CTTGTCAAAGACAAG Minus from sperm

C At the second replication the new T's pair with A

Alkylated G's on Plus are fixed as A's:

AAACAATTTCTGTTC New-New Plus (note new A's)
TTTGTTAAAGACAAG New Minus

C's paired with alkylated G's on Minus fixed as T's
GAACAGTTTTTGTTT New Plus
CTTGTCAAAAACAAA New-New Minus (note new A's)

Figure 3 (A) Alkylated G's as a result of EMS treatment affecting the DNA of haploid gametes are shown in red. (B) During the first round of replication, the alkylated G's will be mis-paired with T's. (C) In the next round of replication, the new T's will pair with A's. Either one of the chromosomes shown in C could enter the P1 cell and give rise to the germ line and the next generation of offspring. If the chromosome giving rise to the germ line comes from the "New Minus" strand, the mutations will appear as G→A changes when compared to the reference sequence. If the chromosome giving rise to the germ line comes from the "New Plus" strand, the mutations will appear as a C→T changes. Other segregant possibilities are not shown for simplicity. Note that all the mutational changes along the chromosome are of the same type, either G→A or C→T, and that the type will be passed on to the progeny in the next generation.

previous published description of *e61* (Thacker *et al.* 2006). The other marker, *unc-13 (e450)*, is situated outside the duplicated region, and thus we expected that 90–100% of the reads would correspond to the *e450* allele. We observed that 96% (55/56 reads) had a T at position 7,435,169 in a gene encoding *unc-13* (Ahmed *et al.* 1992), whereas in our N2 strain there is a C at this position. The change would result in a STOP codon replacing the normal glutamine (Q) residue in the 13th exon. Previously, *e450* was known genetically to introduce a stop codon into *unc-13* (Waterston 1981); however, this is the first report of the nucleotide change responsible for this allele.

Having demonstrated that we could correctly identify SNVs corresponding to known mutations, we set out to find the molecular lesion for *let-504*. The *let-504* gene was mapped previously to the *hDf7* region (Johnsen *et al.* 2000). Genetic mapping data placed the left breakpoint of *hDf7* to the right of *unc-89*, and the right breakpoint to the left of *anc-1*, thus positioning *hDf7* between 4.10 and 4.32 Mbp on chromosome I (Figure 1). We used both manual and bioinformatics analysis looking for SNVs within the *hDf7* region. Similar to the *dpy-5 (e61)* mutation, we expected the SNV to have an allelic ratio close to 70–80%, and we discounted any SNV that was present in all the reads. For example, we found a single base-pair deletion in R12E2.1 that was present in all the reads, and we discounted it as a candidate mutation.

We identified three SNVs that satisfied our criteria (Table 2). A G→A change with 70% read support (35/50) was found in a noncoding region 18 bp upstream of H31G24.3. No other SNVs were found in H31G24.3. Sanger sequencing revealed that the same mutation occurs upstream of H31G24.3 in all the *let-504* alleles. Since the different alleles of *let-504* have different arrest stages, this mutation is likely not the *h448* mutation. A C→T mutation was found in the last intron of E01A2.1 and was supported by 75% of the reads (111/148). This mutation did not disrupt the coding

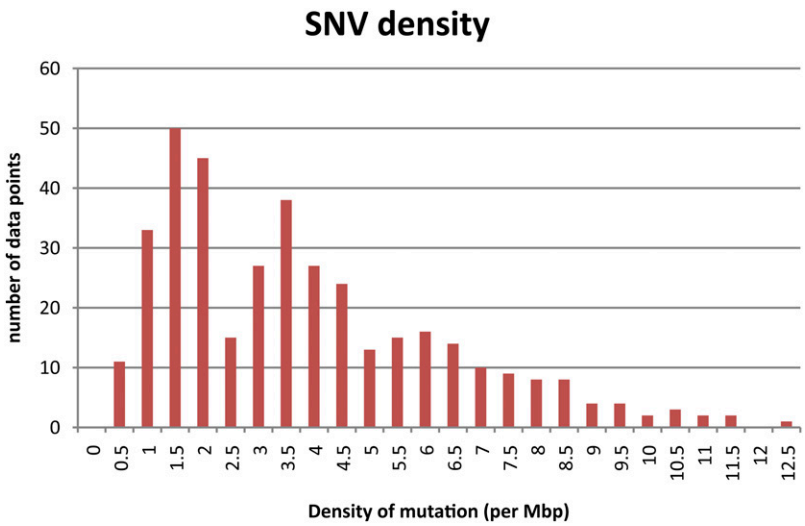


Figure 4 The number of EMS-induced changes per 1 Mbp. This plot was generated by combining the rates from KR772 (excluding chromosome I), RB5002, VC1923, and VC1924. Two prominent peaks are clearly observed: one at 1.5/Mbp and another at 3.5/Mbp. A smaller peak was also observed at 6/Mbp. The data for RB5002, VC1923, and VC1924 are from Flibotte *et al.* (2010).

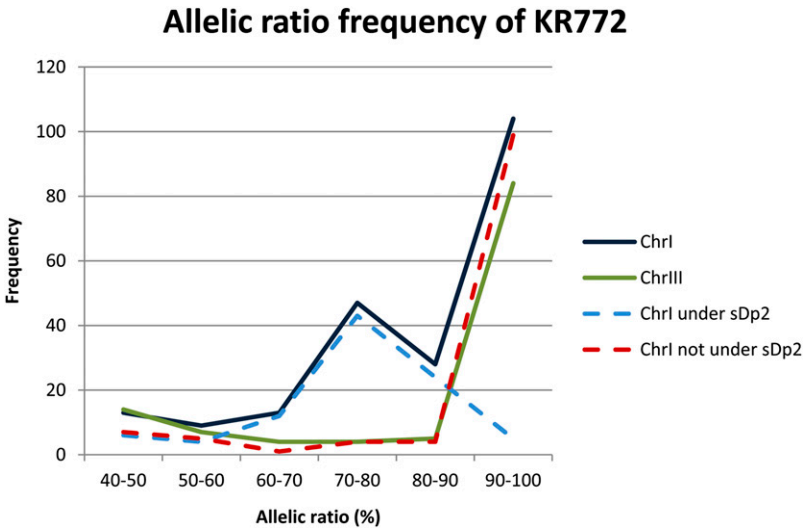


Figure 5 Allelic ratio in KR772 for the whole chromosome I, whole chromosome III, part of chromosome I under *sDp2*, and part of chromosome I not under *sDp2*. Allelic ratio is presented as the percentage of reads that show SNV at a particular nucleotide position. In the *sDp2* region, the peak at 70–80% represents mutations homozygous in the homologs with a wild-type allele in *sDp2*.

region nor any splice signals and is thus unlikely to be the cause of *h448*. The third mutation was in E01A2.4 and caused a C→T change at position 4,132,191 in 67% of the reads (42/63). This SNV changed the third nucleotide of the first codon from ATG (methionine) to ATA (isoleucine) and effectively removed the start codon for E01A2.4. Given the allelic ratio (67%) and the fact that it is a C→T change, we propose that E01A2.4 is *let-504*.

Identification of let-504 as E01A2.4, a NFκB-activating protein ortholog

We PCR-amplified and sequenced the coding region of E01A2.4 from strains carrying the remaining four *let-504* alleles (Table 1). Sanger sequencing revealed nonsense mutations in E01A2.4 for three of the alleles: *h137*, *h844*, and *h888* (Figure 7). *h888* changes the 41st codon from TGG (W) to TGA (STOP); *h844* changes the 151st codon from CAA (Q) to TAA (STOP); and *h137* changes the 358th codon from CAG (Q) to TAG (STOP). The fourth allele (*h327*), which was generated with gamma radiation, did not contain a SNV in the coding region of E01A2.4 and may be a complex mutation not detectable by PCR amplification and sequencing.

The lethal phenotypes of *let-504* alleles correlate with the positions of the alleles in E01A2.4. Using gonadal indexing, we determined that *h888* and *h844*, which remove most of the protein sequence, have the most severe phenotypes, arresting at the mid-larval stage (Figure S4). The milder phenotype, that of sterile adults, was seen in *h137* and *h448* (Figure S4). The *h137* allele truncates the protein close to the end whereas the *h448* allele has a mutated start codon. The milder phenotype of *h137* may result from readthrough of the stop codon, and in the case of *h448* might be due to the use of an alternative start codon that allows a truncated protein to be made. However, we observed that the *tm4719* allele, which removes amino acids 257–404 in E01A2.4, also produces sterile adult animals. Thus, it is possible that the first 256 amino acids contain information required for progression beyond the mid-larval stage.

To further confirm that *let-504* is E01A2.4, we carried out transgenic fosmid rescue and complementation test experiments. The transgenic rescue experiment was done by crossing *let-504 (h448)* animals to a strain carrying the transgenic fosmid WRM0614bH01, which contains E01A2.4. However, we were not able to observe rescue. This could be for many

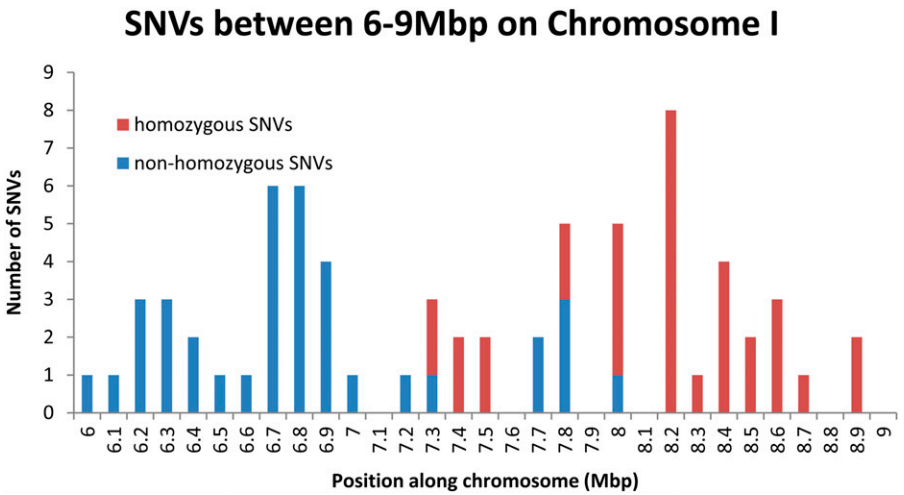


Figure 6 Chromosomal I region between 6 and 9 Mbp. The blue bars represent nonhomozygous SNVs, and the red bars represent homozygous SNVs. An SNV with an allelic ratio between 40 and 89% is considered as nonhomozygous. An SNV with allelic ratio ≥90% is considered as homozygous. A nonhomozygous mutation first occurred at 7.3 Mbp.

Table 2 Candidate *let-504* mutations

Chromosome	Position	Reference	Mutation	Reference read support	Mutation read support	Allelic ratio (%)	Genomic environment
I	4116834	G	A	15	35	70	18 bp upstream of H31G24.3
I	4132191	C	T	20	41	67.21	First exon of E01A2.4
I	4142061	C	T	37	111	75	Last intron of E01A2.1

reasons. Because of the sterility phenotype and morphological defects in the gonad, it is possible that E01A2.4 expression is required in the germ line, where expression of transgenic fosmid may be suppressed. In the absence of a rescuing construct, we carried out a complementation test with the *tm4719* allele. We predicted that worms carrying both *h448* and *tm4719* alleles would arrest as sterile adults if *let-504* is E01A2.4. Thus, we constructed *h448/tm4719* animals (see *Materials and Methods*) and examined 83 sterile adults from *h448* mothers, which also carried the *tm4719* allele. Thus, *h448* failed to complement *tm4719*. We conclude that E01A2.4 is the coding region for the essential gene *let-504*.

BLAST searches with the E01A2.4 protein sequence show sequence similarity to the human gene NFκB-activating protein (NKAP) with 30% identity at the protein level. In humans, NKAP is a transcriptional repressor that associates with the NOTCH corepressor complex and is required for T-cell development (Pajerowski *et al.* 2009).

Concluding remarks

The issue of heterozygosity continues to be a challenge in genome sequence analysis. Here, we have demonstrated that heterozygous SNVs can be identified effectively using information from the mutational landscape and allelic ratios. An application of our approach is the identification of lethal mutations in essential genes. About 3000–5000 of the 20,000 genes in *C. elegans* are estimated to be essential for development and survival. Over the past 25 years, thousands of lethal alleles corresponding to >500 essential genes have been isolated. These mutations are maintained as heterozygous mutations using translocations and duplications. In the case of the duplication *sDp2*, 237 essential genes have been mapped genetically. Correlating coding regions to these lethal mutations has been slow and laborious. We have shown that WGS is a time-efficient and cost-effective way for further characterizing essential genes. Our approach is also applicable to situations in which the heterozygous mutations exist in a 1:1 allelic ratio (*m*/+). With deep enough coverage, a 50% SNV allelic ratio will stand out against the statistical noise of the sequencing methodology. The coverage required will depend upon the read length and the sequencing methodology and can be calculated for specific situations. In addition, we have provided a better molecular understanding of EMS mutation fixation, which may be a useful tool for identifying alkylating agent-induced lesions in *C. elegans* and other model organisms. The fact that a particular type of EMS mutation occurs in contiguous blocks reduces the number of non-informative changes and provides prediction with regard to mutation type. The approach taken here is readily applicable to the

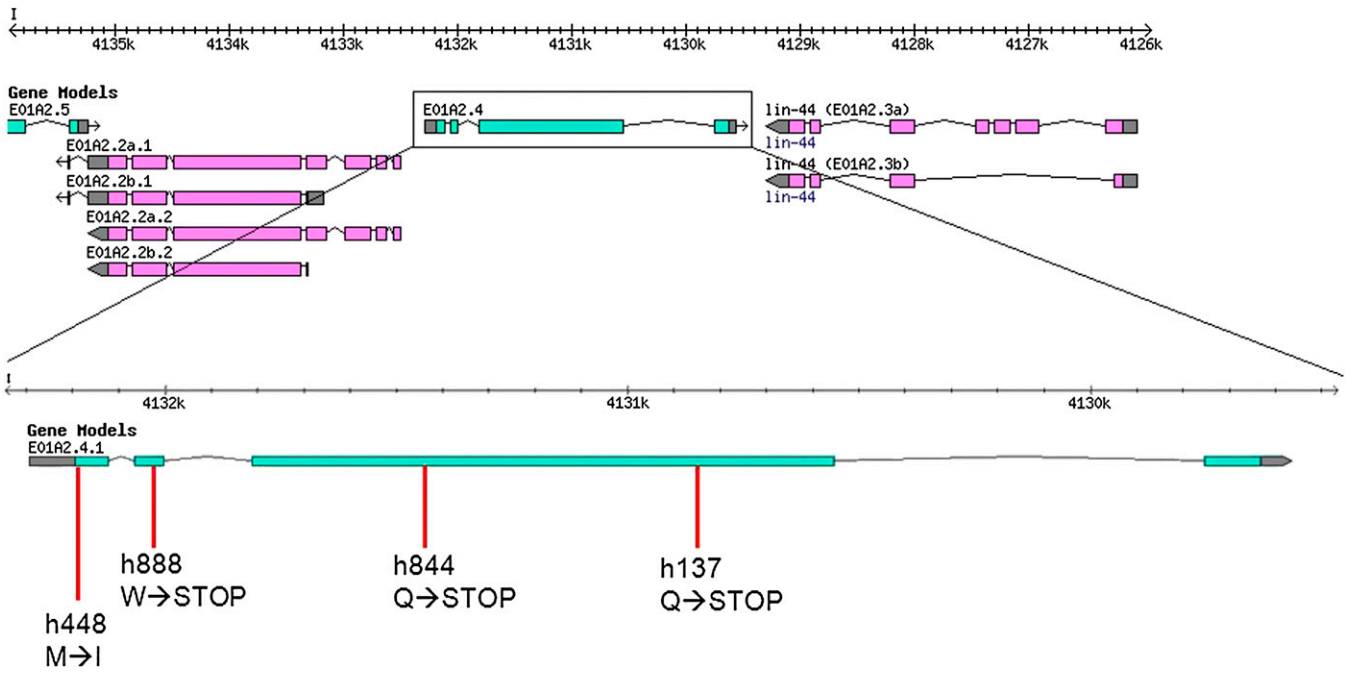


Figure 7 Location of *let-504* alleles. The changes underneath the allele name indicate amino acid changes.

rest of the lethal collection as well as to other phenotypes. As essential genes often encode highly conserved proteins that act either in cell maintenance or in a developmentally critical pathway, identification of the coding regions corresponding to the mutant collection will greatly increase both available genetic resources and information about gene function. Identification of the molecular basis of these genes is of value for both our understanding of animal biology and the study of human disease.

Acknowledgments

The authors thank Martin Jones and Jason Luce for their input and help with the manuscript. A special thanks to S. Mitani for *tm4719*. This work has been supported by grants from the Canadian Institutes of Health Research to A.M.R. and the Natural Sciences and Engineering Council to D.L.B.

Literature Cited

Ahmed, S., I. N. Maruyama, R. Kozma, J. Lee, S. Brenner *et al.*, 1992 The Caenorhabditis elegans unc-13 gene product is a phospholipid-dependent high-affinity phorbol ester receptor. Biochem. J. 287(Pt. 3): 995–999.

Bautz, E., and E. Freese, 1960 On the mutagenic effect of alkylating agents. Proc. Natl. Acad. Sci. USA 46: 1585–1594.

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

Cronn, R., A. Liston, M. Parks, D. S. Gernandt, R. Shen *et al.*, 2008 Multiplex sequencing of plant chloroplast genomes using Solexa sequencing-by-synthesis technology. Nucleic Acids Res. 36: e122.

Doitsidou, M., R. J. Poole, S. Sarin, H. Bigelow, and O. Hobert, 2010 C. elegans mutant identification with a one-step whole-genome-sequencing and SNP mapping strategy. PLoS ONE 5: e15435.

Edgley, M. L., D. L. Bailey, D. L. Riddle, and A. M. Rose, 2006 Genetic balancers (April 6, 2006), WormBook, ed. The C. elegans Research Community, WormBook, doi/10.1895/wormbook.1.89.1, http://www.wormbook.org.

Flibotte, S., M. L. Edgley, I. Chaudhry, J. Taylor, S. E. Neil *et al.*, 2010 Whole-genome profiling of mutagenesis in *Caenorhabditis elegans*. Genetics 185: 431–441.

Fraser, A. G., R. S. Kamath, P. Zipperlen, M. Martinez-Campos, M. Sohrmann *et al.*, 2000 Functional genomic analysis of C. elegans chromosome I by systematic RNA interference. Nature 408: 325–330.

Greene, E. A., C. A. Codomo, N. E. Taylor, J. G. Henikoff, B. J. Till *et al.*, 2003 Spectrum of chemically induced mutations from a large-scale reverse-genetic screen in Arabidopsis. Genetics 164: 731–740.

Gresham, D., D. M. Ruderfer, S. C. Pratt, J. Schacherer, M. J. Dunham *et al.*, 2006 Genome-wide detection of polymorphisms at nucleotide resolution with a single DNA microarray. Science 311: 1932–1936.

Hillier, L. W., G. T. Marth, A. R. Quinlan, D. Dooling, G. Fewell *et al.*, 2008 Whole-genome sequencing and variant discovery in C. elegans. Nat. Methods 5: 183–188.

Hobert, O., 2010 The impact of whole genome sequencing on model system genetics: get ready for the ride. Genetics 184: 317–319.

Howell, A. M., and A. M. Rose, 1990 Essential genes in the hDf6 region of chromosome I in *Caenorhabditis elegans*. Genetics 126: 583–592.

Howell, A. M., S. G. Gilmour, R. A. Mancebo, and A. M. Rose, 1987 Genetic analysis of a large autosomal region in Caenorhabditis elegans by the use of a free duplication. Genet. Res. 49: 207–213.

Janke, D. L., J. E. Schein, T. Ha, N. W. Franz, N. J. O'Neil *et al.*, 1997 Interpreting a sequenced genome: toward a cosmid transgenic library of Caenorhabditis elegans. Genome Res. 7: 974–985.

Johnsen, R. C., S. J. Jones, and A. M. Rose, 2000 Mutational accessibility of essential genes on chromosome I(left) in Caenorhabditis elegans. Mol. Gen. Genet. 263: 239–252.

Koboldt, D. C., K. Chen, T. Wylie, D. E. Larson, M. D. McLellan *et al.*, 2009 VarScan: variant detection in massively parallel sequencing of individual and pooled samples. Bioinformatics 25: 2283–2285.

Lewis, E. B., and F. Bacher, 1968 Method of feeding ethyl methanesulfonate (EMS) to Drosophila males. Drosoph. Inf. Serv. 43: 193.

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

Maydan, J. S., H. M. Okada, S. Flibotte, M. L. Edgley, and D. G. Moerman, 2009 De novo identification of single nucleotide mutations in *Caenorhabditis elegans* using array comparative genomic hybridization. Genetics 181: 1673–1677.

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.

Pajerowski, A. G., C. Nguyen, H. Aghajanian, M. J. Shapiro, and V. S. Shapiro, 2009 NKAP is a transcriptional repressor of notch signaling and is required for T cell development. Immunity 30: 696–707.

Robinson, J. T., H. Thorvaldsdottir, W. Winckler, M. Guttman, E. S. Lander *et al.*, 2011 Integrative genomics viewer. Nat. Biotechnol. 29: 24–26.

Rose, A. M., D. L. Baillie, and J. Curran, 1984 Meiotic pairing behavior of two free duplications of linkage group I in Caenorhabditis elegans. Mol. Gen. Genet. 195: 52–56.

Rose, A. M., N. J. O'Neil, M. Bilenky, Y. S. Butterfield, N. Malhis *et al.*, 2010 Genomic sequence of a mutant strain of Caenorhabditis elegans with an altered recombination pattern. BMC Genomics 11: 131.

Rosenbluth, R. E., and D. L. Baillie, 1981 The genetic analysis of a reciprocal translocation, eT1(III; V), in *Caenorhabditis elegans*. Genetics 99: 415–428.

Sarin, S., S. Prabhu, M. M. O'Meara, I. Pe'er, and O. Hobert, 2008 Caenorhabditis elegans mutant allele identification by whole-genome sequencing. Nat. Methods 5: 865–867.

Sarin, S., V. Bertrand, H. Bigelow, A. Boyanov, M. Doitsidou *et al.*, 2010 Analysis of multiple ethyl methanesulfonate-mutagenized *Caenorhabditis elegans* strains by whole-genome sequencing. Genetics 185: 417–430.

Shen, Y., S. Sarin, Y. Liu, O. Hobert, and I. Pe'er, 2008 Comparing platforms for C. elegans mutant identification using high-throughput whole-genome sequencing. PLoS ONE 3: e4012.

Simms, C. L., and D. L. Baillie, 2010 A strawberry notch homolog, let-765/nsh-1, positively regulates lin-3/egf expression to promote RAS-dependent vulval induction in C. elegans. Dev. Biol. 341: 472–485.

Sonnichsen, B., L. B. Koski, A. Walsh, P. Marschall, B. Neumann *et al.*, 2005 Full-genome RNAi profiling of early embryogenesis in Caenorhabditis elegans. Nature 434: 462–469.

Sulston, J., and J. Hodgkin, 1988 Methods, pp. 587–606 in *The Nematode Caenorhabditis elegans*, edited by W. Wood. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Thacker, C., J. A. Sheps, and A. M. Rose, 2006 Caenorhabditis elegans dpy-5 is a cuticle procollagen processed by a proprotein convertase. Cell. Mol. Life Sci. 63: 1193–1204.

Waterston, R. H., 1981 A second informational suppressor, SUP-7 X, in *Caenorhabditis elegans*. Genetics 97: 307–325.

Communicating editor: D. I. Greenstein

GENETICS

Supporting Information
http://www.genetics.org/content/suppl/2012/01/20/genetics.111.137208.DC1

Allelic Ratios and the Mutational Landscape Reveal Biologically Significant Heterozygous SNVs

Jeffrey S.-C. Chu, Robert C. Johnsen, Shu Yi Chua, Domena Tu, Mark Dennison, Marco Marra, Steven J. M. Jones, David L. Baillie, and Ann M. Rose

DOI: 10.1534/genetics.111.137208

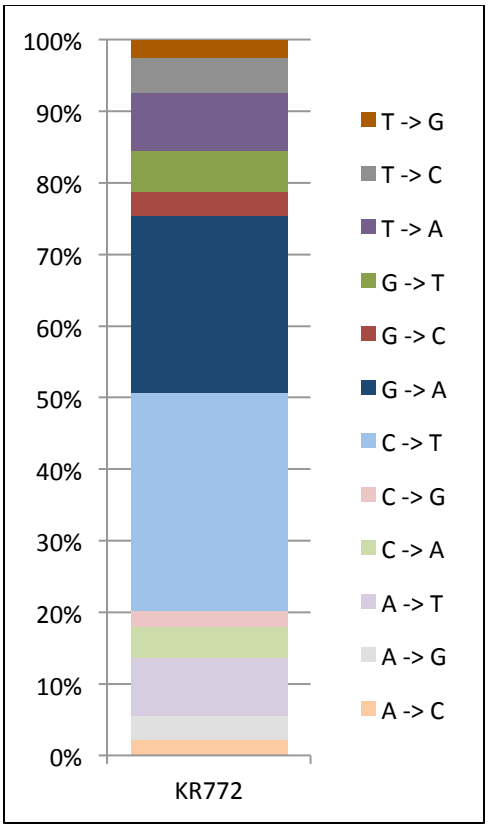


Figure S1 The distribution of single nucleotide variations (SNVs) in KR772.

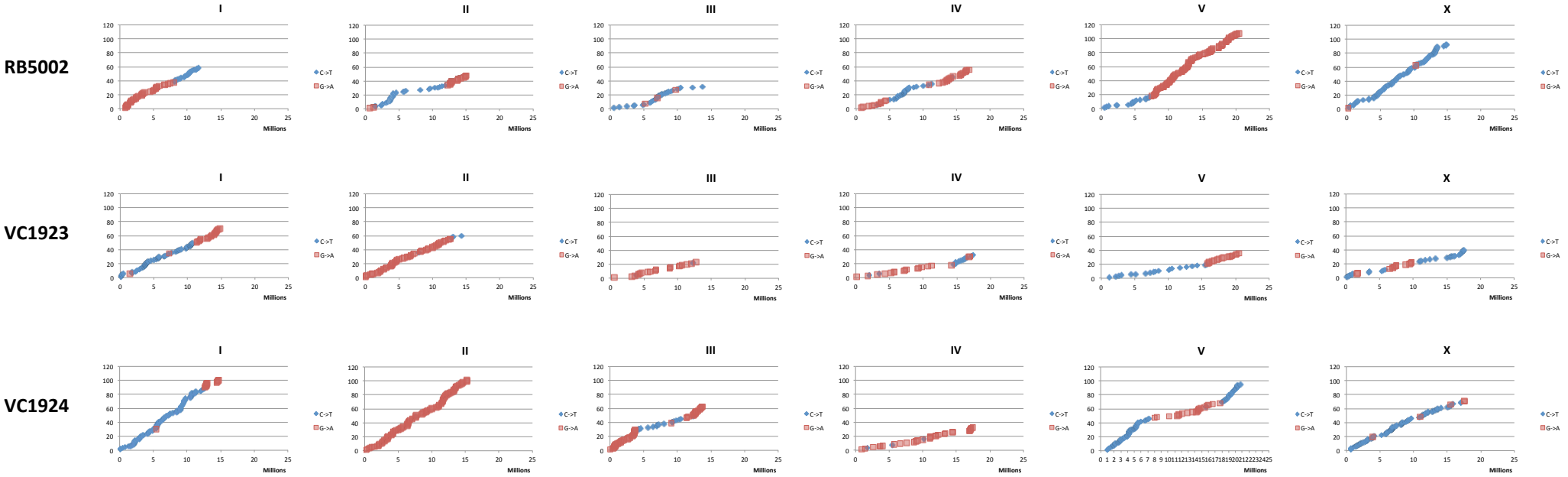
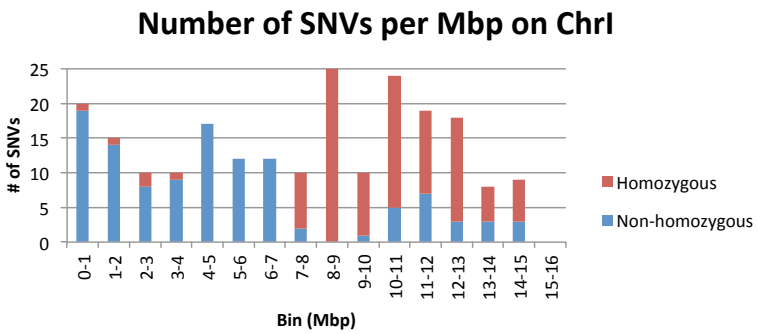
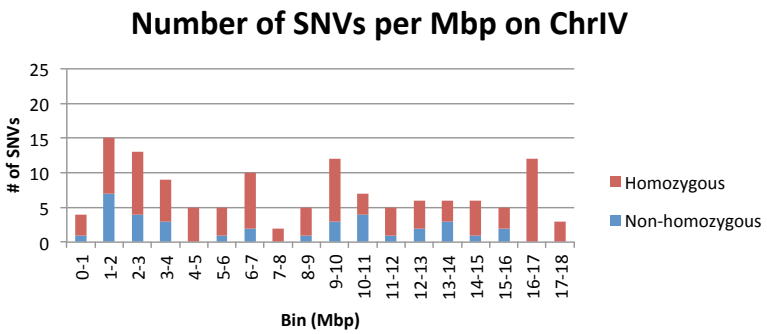
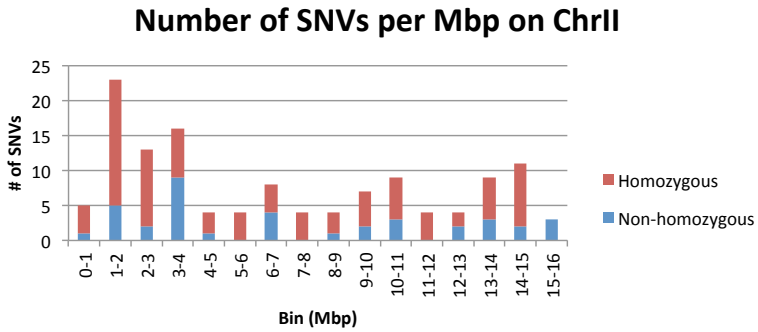
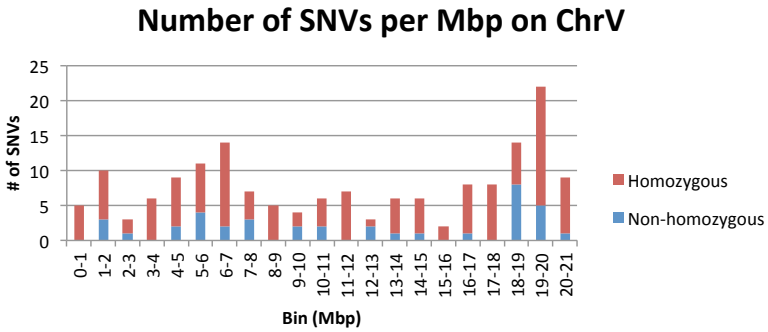


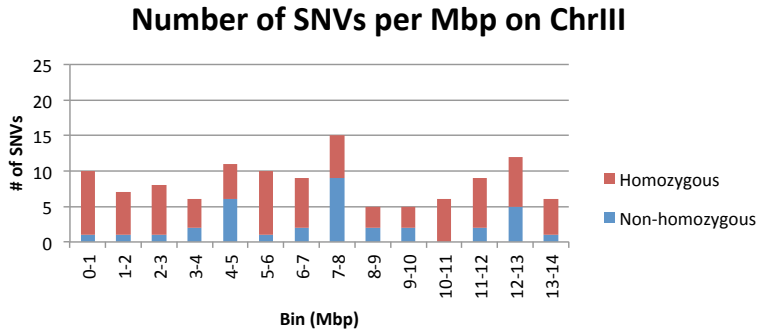
Figure S2 Non-random distribution of EMS induced changes are seen in RB5002, VC1923, and VC1924. The X-axis represents the length of chromosomes. The Y-axis indicates each SNV ID. The data used to generate this figure was from the supplementary table of (FLIBOTTE *et al.* 2010).


Number of SNVs per Mbp on ChrI
# of SNVs
Bin (Mbp)
Homozygous
Non-homozygous


Number of SNVs per Mbp on ChrIV
# of SNVs
Bin (Mbp)
Homozygous
Non-homozygous


Number of SNVs per Mbp on ChrII
# of SNVs
Bin (Mbp)
Homozygous
Non-homozygous


Number of SNVs per Mbp on ChrV
# of SNVs
Bin (Mbp)
Homozygous
Non-homozygous


Number of SNVs per Mbp on ChrIII
# of SNVs
Bin (Mbp)
Homozygous
Non-homozygous

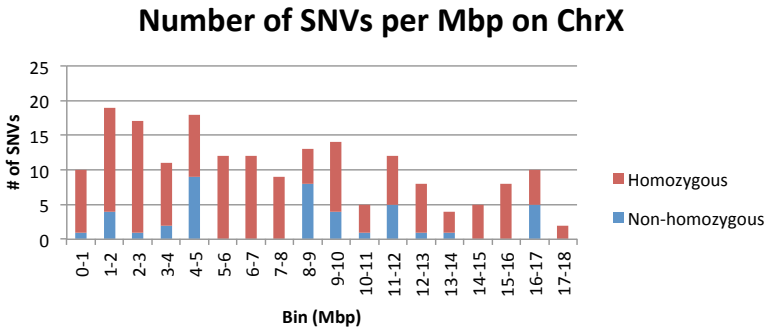

Number of SNVs per Mbp on ChrX
# of SNVs
Bin (Mbp)
Homozygous
Non-homozygous

Figure S3 Distribution of homozygous and non-homozygous SNVs per 1Mbp window in each chromosome. The blue bars represent non-homozygous SNVs and the red bars represent homozygous SNVs. A SNV with allelic ratio between 40% and 89% are considered as non-homozygous. A SNV with allelic ratio 90% or above are considered as homozygous.

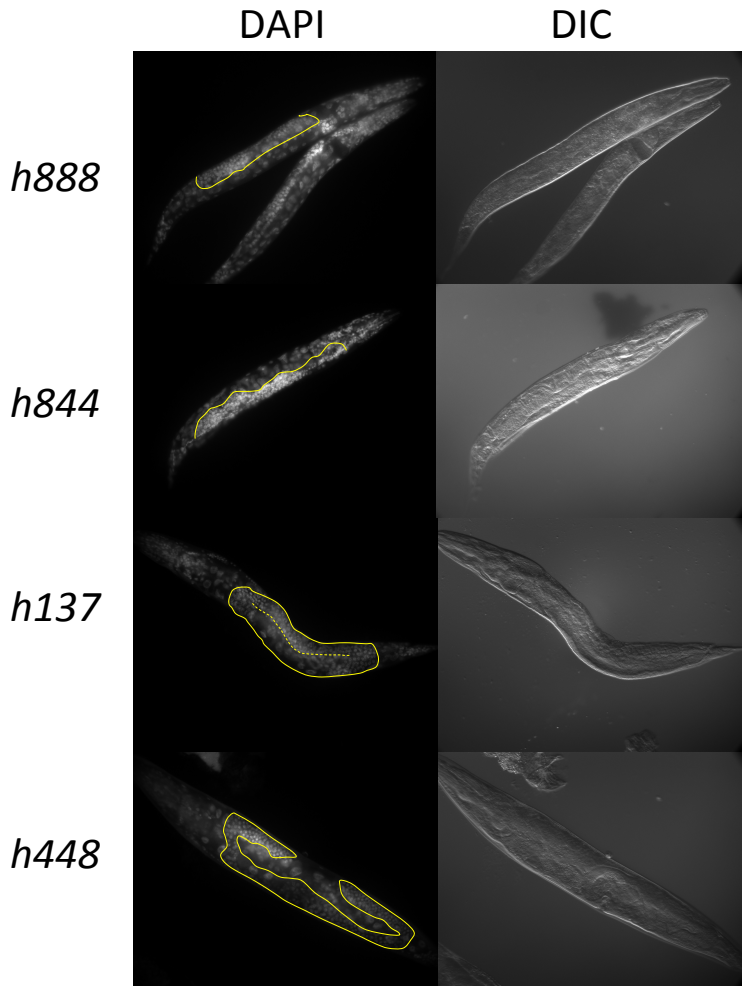


Figure S4 DAPI staining and DIC images for *h888, h844, h137*, and *h448*. The yellow line outlines the gonad. The *h888* and *h844* alleles show the more severe phenotype where the gonadal arms have yet to turn. The *h137* and *h448* alleles show milder phenotype where the gonadal arms have fully turned.

Table S1 All SNVs present in KR772 with more than 20% read support.

Table S1 is available for download at http://www.genetics.org/content/suppl/2012/01/20/genetics.111.137208.DC1 as an excel file.