Worm 2:3, e25081; July/August/September 2013; © 2013 Landes Bioscience

Deep sequencing strategies for mapping and identifying mutations from genetic screens

Steven Zuryn and Sophie Jarriault*

Institut de Génétique et de Biologie Moléculaire et Cellulaire (IGBMC); Institut National de la Santé et de la Recherche Médicale (INSERM) U964/Centre National de la Recherche Scientifique (CNRS) UMR 1704/Université de Strasbourg; Strasbourg, France

Keywords: deep sequencing, whole genome sequencing, WGS, *C. elegans*, forward genetics, next-generation sequencing, NGS, genetic mapping, SNP mapping, EMS

Abbreviations: DNA, deoxyribonucleic acid; RNAi, ribonucleic acid interference; SNP, single nucleotide polymorphism; F_2, filial generation 2; F_1, filial generation 1; RNA, ribonucleic acid; NGS, next-generation sequencing; Mb, megabase; EMS, ethyl methanesulfonate; RAD, restriction site-associated DNA

The development of next-generation sequencing technologies has enabled rapid and cost effective whole genome sequencing. This technology has allowed researchers to shortcut time-consuming and laborious methods used to identify nucleotide mutations in forward genetic screens in model organisms. However, causal mutations must still be mapped to a region of the genome so as to aid in their identification. This can be achieved simultaneously with deep sequencing through various methods. Here we discuss alternative deep sequencing strategies for simultaneously mapping and identifying causal mutations in *Caenorhabditis elegans* from mutagenesis screens. Focusing on practical considerations, such as the particular mutant phenotype obtained, this review aims to aid the reader in choosing which strategy to adopt to successfully clone their mutant.

Introduction

Forward genetics is a powerful means to identify the genes and molecular mechanisms that determine a phenotype. A chemical mutagenesis screen approach, usually employing an agent such as ethyl methanesulfonate (EMS),[1] has certain advantages, such as its unbiased nature in pinpointing DNA elements that are important for a certain phenotype. A mutagenesis screen will also produce valuable mutants, typically stable and often in multiple distinct alleles, which can be utilized as a resource for further characterization of the gene. Indeed, the localization and amino acid changes induced by mutagenesis can itself help to illuminate the role of the gene if, for instance, a functional domain of the corresponding protein is perturbed. Until recently, the main disadvantage of conducting mutagenesis screens was the relative difficulty and time involved in identifying the gene affected by the causal mutation. Although this bottleneck is avoided when resorting to large-scale RNA interference (RNAi) screens,[2] no mutant alleles are obtained, a disadvantage when further analyzing gene function.

Traditionally, in order to identify the causal mutation, single nucleotide polymorphism (SNP) mapping is employed to narrow down a genomic region harboring the DNA lesion.[3-5] Typically, a cross is set up between mutant isolates and an alternative strain of the same species that contains polymorphic nucleotides. A quarter of the resulting second filial generation (F_2) progeny from an F_1 cross or self-cross (depending on the species being studied) will inherit the causal mutation in homozygous form. Unless the mutation is either dominant, or a parental RNA or protein contribution from a heterozygote parent is sufficient to rescue mutant offspring, these F_2 mutant progeny will display the mutant phenotype and can be selected. As a result of Mendelian inheritance of chromatids, as well as mostly random chromosomal recombination during meiosis, the polymorphisms from the parental and mapping strains will be distributed in a roughly 50/50 ratio in the selected F_2, except for those SNPs that are genetically linked to the causal mutation. These linked polymorphisms, which are within relatively close physical proximity to the causal mutation, are statistically less likely to be included in a chromosomal recombination event compared with those further away on the chromosome. Thus, the location of the causal mutation is betrayed by a stretch of the genome marked by a disproportionately high frequency of parental polymorphisms and an underrepresentation of mapping strain polymorphisms. This fundamental rule of chromosomal recombination is the basis of genetic mapping.

Over the past 5 y, methods based on next-generation sequencing (NGS) technology[6] have allowed advances in several research fields leading to the identification of the mutational processes involved in various types of cancers and genetic diseases in non-genetic model organisms such as humans;[7,8] or to obtain insights into phylogenetics and evolutionary processes.[9] Here, we will highlight how this technology has aided in the identification of mutations from forward genetic screens, with a focus on *C. elegans*. Two general approaches to identify causal mutations using NGS have been employed. In a first approach, the causal mutation is mapped to a genomic region using traditional SNP-based

*Correspondence to: Sophie Jariault; Email: sophie@igbmc.fr
Submitted: 04/05/13; Revised: 05/16/13; Accepted: 05/17/13
Citation: http://dx.doi.org/10.4161/worm.25081

approaches prior to sequencing the whole genome,[10] which is used to detect candidate nucleotide mutations within this interval. The second approach involves mapping as well as identifying the causal mutation simultaneously by whole genome sequencing. In this technical-based review, we focus on the alternative strategies developed for this second approach whereby NGS is employed to simultaneously pinpoint chromosomal recombination events (mapping) and detect candidate mutations in this mapped region. These powerful techniques have dramatically sped up a process that was traditionally achieved through detection of nucleotide variants one at a time, usually with the use of restriction endonucleases. It may be intuitive to ask that if the entire genome of a mutant is sequenced, why would there be a need to map a mutation to a specific region of the genome? Deep sequencing of a *C. elegans* genome will reveal the presence of many nucleotide variants that exist in the backgrounds of different strains used across the *C. elegans* community.[11-15] New nucleotide variants, including those within genes, also appear relatively rapidly in strains kept in the laboratory.[16] These widespread variants severely hamper the identification of the true causal mutation(s), thus making the need for mapping essential. Indeed, it is not only in *C. elegans* that background nucleotide variants cause problems in causal mutation identification. This phenomenon is conserved across plants and animals, leading to the establishment of different strategies and methodological tools to simultaneously map and deep sequence mutants of a wide variety of species including Zebrafish,[17] Arabidopsis,[18] Rice,[19] mouse[20] and of course *C. elegans*.[12,21]

SNP-Based Deep Sequence Mapping

First proposed by Lister and colleagues and then tested in *Arabidopsis thaliana*, the SNP-based method of simultaneous mapping and deep sequencing was successful in a proof-of-principle experiment to identify a recessive allele that caused slow growth and light green leaves.[18,22] The principle of the method is the same as classical SNP-based mapping. The difference being that instead of performing PCR, restriction digestion and agarose gel electrophoresis to identify polymorphism frequencies one at a time, deep sequencing is used to identify most of them at once.

To create a mapping population, the mutant strain–produced in a Columbia background–was crossed to the polymorphic Landsberg *erecta* strain. A single genomic DNA sample was prepared from a pool of 500 mutant F_2 plants and an Illumina library was prepared and sequenced to 22-fold genome coverage. Using this method, the authors observed a stretch of DNA devoid of Landsberg polymorphisms on chromosome 4, which contained a serine to asparagine nonsynonymous codon change in the *AT4G35090* gene.

This same technique has been adapted to *C. elegans* using the N2 Bristol parental strain and Hawaiian CB4856[23] as a mapping strain.[21] Essentially, the principle and methodology (**Fig. 1A**) are similar to that used for Arabidopsis. The mutant to be mapped (in an N2 background) is crossed to the CB4856 strain and F_2 recombinant progeny that display the mutant phenotype are singled onto fresh plates and grown for a couple of generations (**Fig. 2**). These independent populations are then pooled and DNA is extracted and sequenced. In a proof-of-principle experiment, Doitsidou et al. mapped the location of a mutant defective in dopaminergic neuron specification to the X chromosome. Encouragingly, a far lesser number of F_2 recombinant progeny were used here than the 500 previously used for Arabidopsis. When 50 F_2 recombinants were selected, grown and pooled, a mapping interval of 2.1 megabases (Mb) was defined on chromosome X.[21] When 20 F_2 recombinants were used, a mapping interval of 4.9 Mb was defined on chromosome X. Within these mapped regions, a premature stop codon was found to be present in the *vab-3* gene, disruption of which was found to control the generation of dopaminergic neurons.

A similar SNP-based mapping approach using deep sequencing was used by O'Rourke et al. (2011) to identify several mutant alleles in *C. elegans*. However, instead of sequencing total genomic DNA from pooled recombinant F_2 of a cross between the N2 mutants and CB4856, deep sequencing was performed only on selectively amplified genomic fragments that were in close proximity to an EcoRI restriction site. SNPs within this restriction site-associated DNA (RAD) were detected and used to map a region of low-density CB4856 polymorphisms in the genome. This RAD mapping approach, adapted from *Drosophila* and other species,[24,25] allows for a higher throughput of mutant mapping as sequencing volume is reduced. As only a fraction of the genome is sequenced, it does not, however, allow simultaneous identification of the causal mutation. Instead, the authors returned to the mutant and pulled down the genomic interval corresponding to the mapped region by annealing sheared mutant genomic DNA to corresponding biotinylated fosmid fragments attached to magnetic beads.[26] These mutant fragments were then deep sequenced to identify candidate causal mutations within.

EMS-Based Deep Sequence Mapping

EMS-based mapping works on a similar yet different principle where instead of using polymorphisms between two different strains to detect a genomic region linked to the causal mutation, the nucleotide changes caused by the EMS mutagenesis itself are used (**Figs. 1B and 3**). The approach was first proposed and demonstrated in a proof-of-principle experiment in *C. elegans* where several independent mutations that cause a defect in a transdifferentiation event, where the rectal epithelial Y cell converts into the motoneuron PDA cell,[27-29] were mapped to different regions on chromosomes III and X.[12] One of these, the previously uncharacterised mutant strain carrying the *fp6* allele was confirmed to be caused by a missense mutation in the hox gene *egl-5*, situated within the middle of the mapped region on chromosome III.[12]

The principle of the method is simple. During mutagenesis treatment, EMS randomly distributes mutations across the genome. EMS has a strong bias for inducing Guanine (G) to Adenosine (A) nucleotide transitions.[14,30,31] The reciprocal strand of DNA presents this mutation as a Cytosine (C) to Thymine (T) substitution. The majority of these EMS-induced nucleotide changes are removed from the background by back/outcrossing

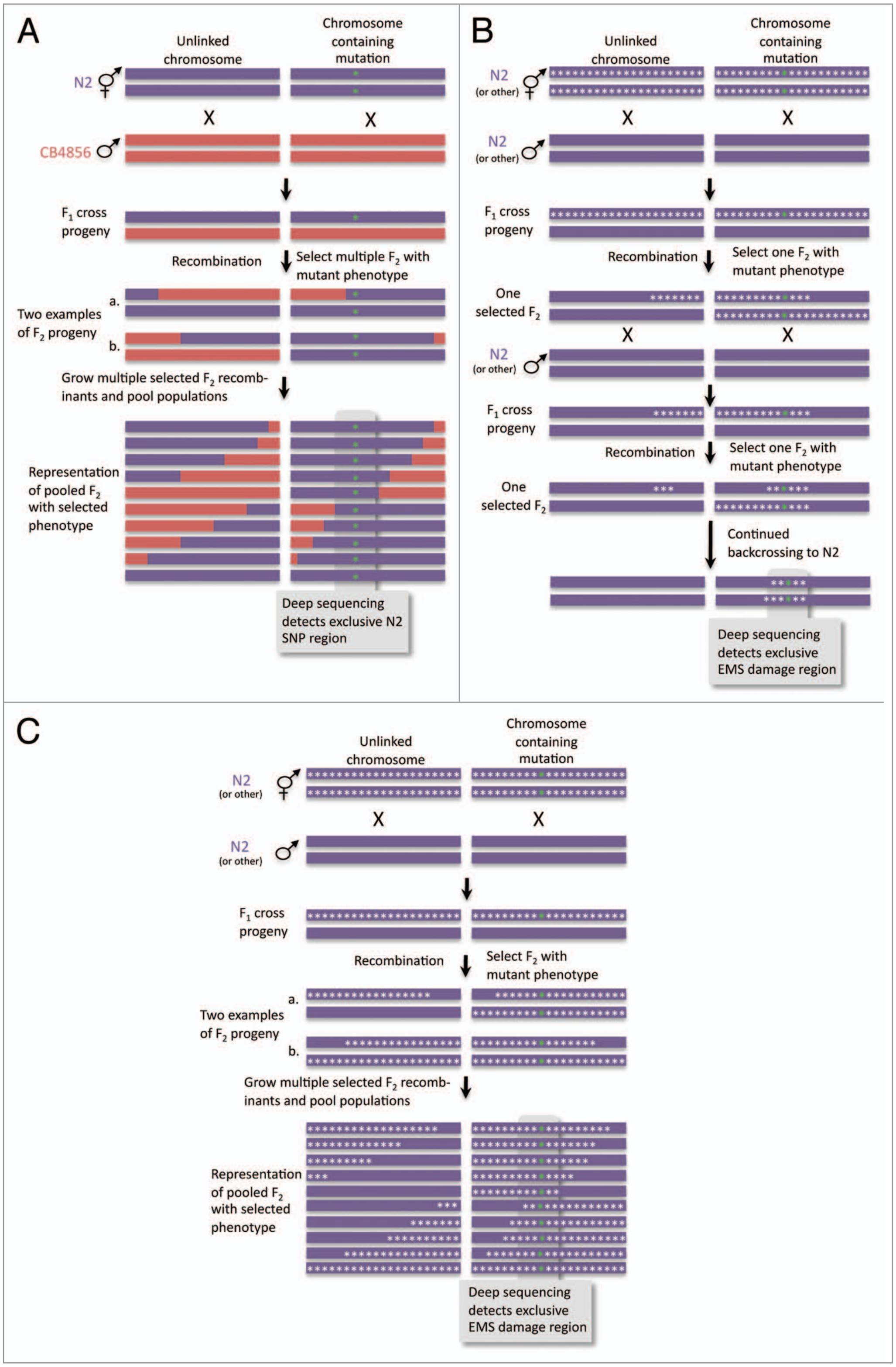

A
Unlinked chromosome
Chromosome containing mutation
N2 ⚥
X
X
CB4856 ♂
F1 cross progeny
Recombination
Select multiple F2 with mutant phenotype
Two examples of F2 progeny
a.
b.
Grow multiple selected F2 recombinants and pool populations
Representation of pooled F2 with selected phenotype
Deep sequencing detects exclusive N2 SNP region
B
Unlinked chromosome
Chromosome containing mutation
N2 (or other) ⚥
X
X
N2 (or other) ♂
F1 cross progeny
Recombination
Select one F2 with mutant phenotype
One selected F2
X
X
N2 (or other) ♂
F1 cross progeny
Recombination
Select one F2 with mutant phenotype
One selected F2
Continued backcrossing to N2
Deep sequencing detects exclusive EMS damage region
C
Unlinked chromosome
Chromosome containing mutation
N2 (or other) ⚥
X
X
N2 (or other) ♂
F1 cross progeny
Recombination
Select F2 with mutant phenotype
Two examples of F2 progeny
a.
b.
Grow multiple selected F2 recombinants and pool populations
Representation of pooled F2 with selected phenotype
Deep sequencing detects exclusive EMS damage region

Figure 1. Principle of each of the simultaneous mapping and mutant identification methods used in *C. elegans*. (**A**) SNP-based strategy requiring crossing of the mutant (N2 background) to CB4856, pooling of multiple isolated F$_2$ recombinants that display the phenotype and deep sequencing to discover a genomic region where only N2 SNPs reside. This region should contain the causal mutation (green star). Adapted from reference 34, (**B**) EMS-based strategy whereby the mutant is backcrossed multiple times until EMS-induced mutations (white stars) are cleared, except for those genetically linked to the causal mutation (green star). The cluster of EMS-induced GA > CT nucleotide changes is detected by deep sequencing and localized to a genomic region that should contain the causal mutation. (**C**) A bulk-segregant approach to the EMS-based mapping strategy that involves a single backcross followed by pooling of F$_2$ recombinants and deep sequencing to identify a genomic region where EMS induced GA > CT are exclusive (white stars). In this region, the causal mutation (green star) should be located.

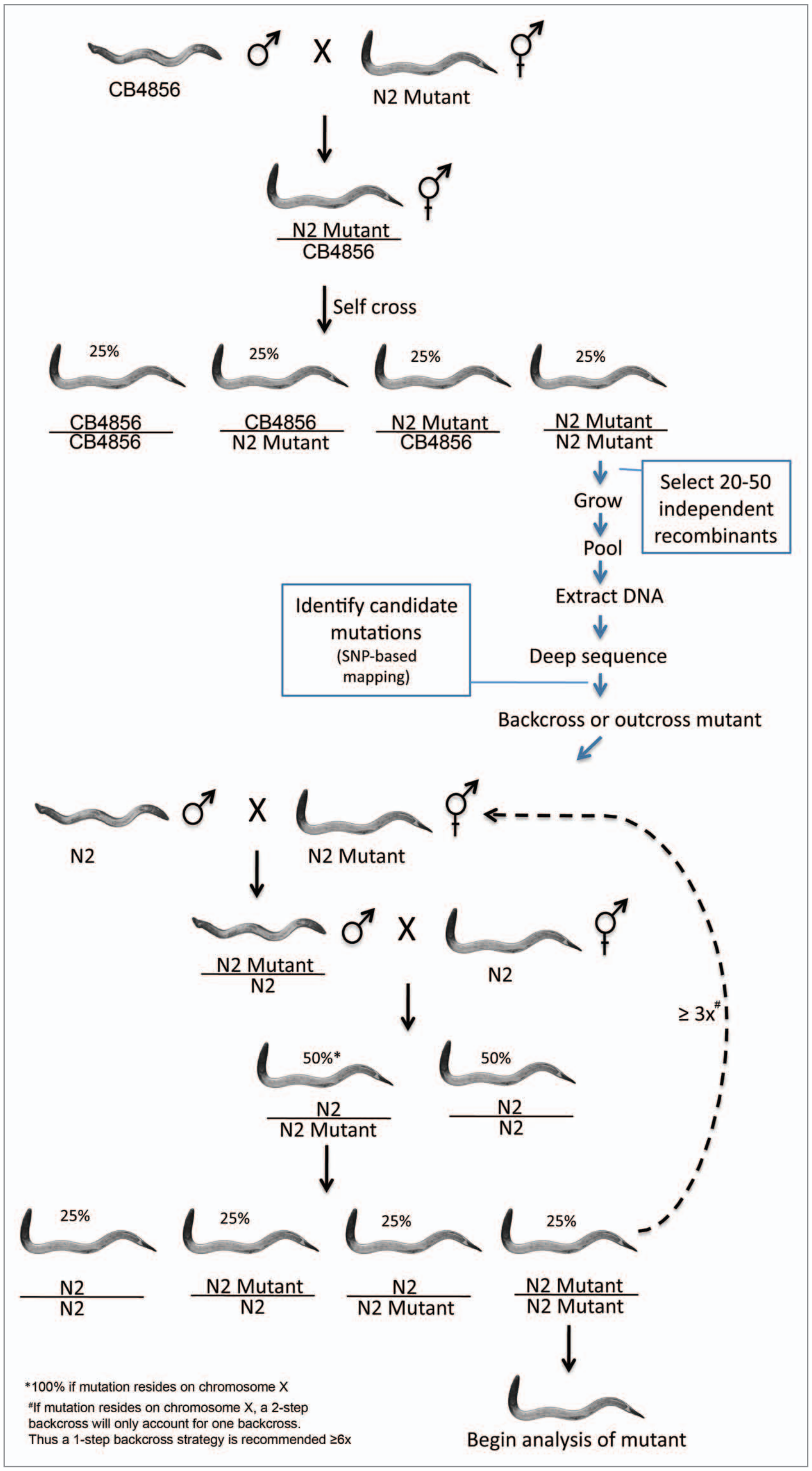

CB4856
♂
X
N2 Mutant
⚥
N2 Mutant / CB4856
Self cross
25%
CB4856 / CB4856
CB4856 / N2 Mutant
N2 Mutant / CB4856
N2 Mutant / N2 Mutant
Select 20-50 independent recombinants
Grow
Pool
Extract DNA
Identify candidate mutations (SNP-based mapping)
Deep sequence
Backcross or outcross mutant
N2
N2 Mutant
N2 Mutant / N2
N2
≥ 3x#
50%*
50%
N2 / N2 Mutant
N2 / N2
N2 / N2
N2 Mutant / N2
N2 / N2 Mutant
N2 Mutant / N2 Mutant
*100% if mutation resides on chromosome X
#If mutation resides on chromosome X, a 2-step backcross will only account for one backcross. Thus a 1-step backcross strategy is recommended ≥6x
Begin analysis of mutant

Figure 2. Flowchart of SNP-based simultaneous mapping and mutant identification using deep sequencing. This strategy is similar to traditional SNP mapping techniques performed before whole genome sequencing was feasible. However, instead of analyzing the linkage of SNPs to a causal mutation one at a time through laborious identification procedures, deep sequencing is used to interrogate most of them simultaneously (see ref. 21). Twenty to 50 F_2 progeny displaying the mutant phenotype are picked from a cross between the mutant (usually in a Bristol N2 background) and a polymorphic strain such as CB4856. They are then singled onto fresh plates, grown and pooled before DNA is extracted and sequenced. From the deep sequencing data and with the use of available software,[34] a genomic region linked to the causal mutation will be identified and nucleotide mutations within this region will also be revealed. A list of candidate causal mutations that affect genes within this region will also be assembled. The original mutant is then subjected to multiple rounds of backcrossing or outcrossing (≥ six simple backcrosses or equivalent recommended) to remove all other EMS-induced mutations from its background and then tested via several methods (e.g., complementation analysis and transgenic rescue) to determine which candidate mutation is causing the phenotype.

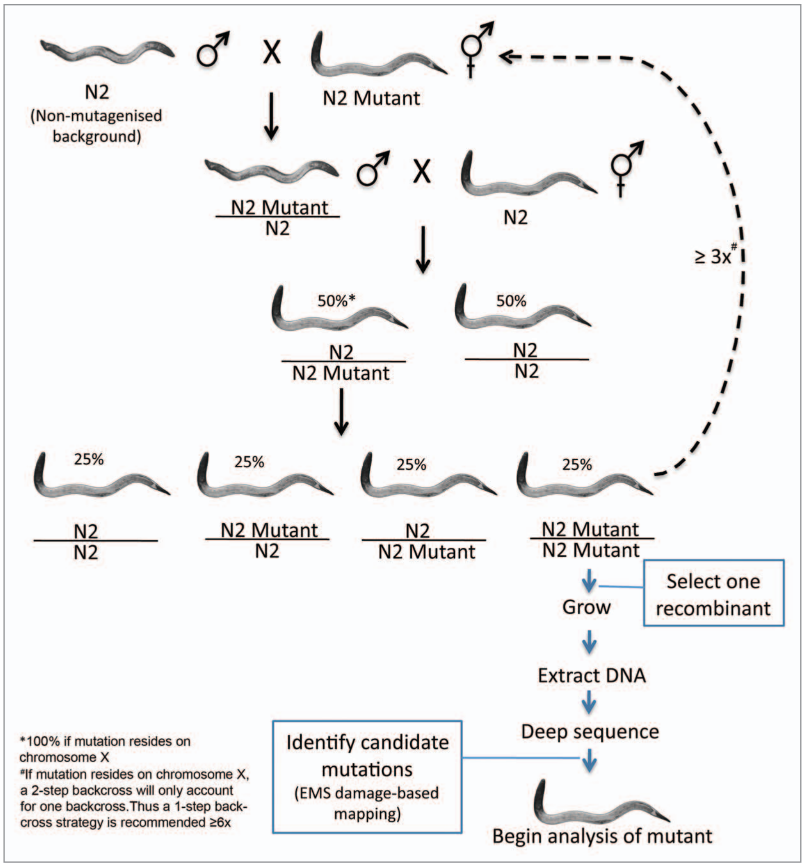
of the mutant to a wild-type strain (for instance the pre-mutagenised strain, or the N2 reference strain). Several (three to six) rounds of backcrossing ensure that enough recombination events between the mutant background and the un-mutated background have occurred so that a distinct "hot spot" of EMS-induced DNA damage can be observed when visualizing the genome sequence (**Fig. 1B**). Thus, these genetically linked EMS-induced nucleotide changes remain present after backcrossing and can be easily observed as G-to-A or C-to-T variants in a genome sequence. We have tested and confirmed that outcrossing to the pre-mutagenized strain or an alternative strain is equally effective at detecting this EMS damage hot spot.

Since EMS damage rather than SNPs are used to map the location of the causal mutation, this method negates the need to create a mapping population between intra-species polymorphic strains. Therefore, the method is theoretically open to any species where such intra-species strains are not available or characterized. The only requirements are that the species is susceptible to mutagenesis by either EMS or *N*-ethyl-*N*-nitrosourea (ENU),[14] an alternative chemical mutagen that also produces canonical nucleotide changes that can be tracked, and that the species can also be back/outcrossed.

Limiting Factors to Mapping Resolution

Despite the fact that the frequency of EMS-induced nucleotide changes during a typical mutagenesis experiment is lower when compared with the frequency of SNPs between N2 and Hawaiian strains, the mapping resolution of each method is limited by chromosomal recombination events instead of polymorphism frequency. In our experience, adult worms exposed to a standard[1] 50 mM dose of EMS over 4 h were found to contain one canonical EMS mutation (G > A or C > T) in every 125,000–143,000 base pairs.[12] Polymorphic nucleotide differences between N2 and CB4856 are distributed across the genome at an average approximate frequency of one in every 1,000 base pairs.[13] Mapping using either SNP- or EMS-based approaches is usually resolved to a genomic region with a width in the millions of base pairs, a limitation resulting from recombination frequency on the chromosome carrying the causal mutation.

In *C. elegans*, chromosomal recombination events tend to be restricted to one per homologous pair in most meiosis.[32] Four backcrosses would thus typically result in four recombination events per chromosome, and six backcrosses would result in six recombination events per chromosome, and so on. In our experience, three to six backcrosses are sufficient to effectively strip chromosomes clean of EMS-exposed DNA. On the chromosome containing the causal mutation, a variable width (but usually around 5 Mb) of chromosomal DNA will still contain EMS damage that is linked to the causal mutation. This variability is due to the random chance that at least one out of three to six recombination events (if three to six backcrosses or outcrosses are performed) will occur within a certain distance of the causal mutation, the probability of which decreases the closer the distance is to the mutation. It also depends upon the location of the causal mutation in relation to recombination hotspots, genomic intervals in which crossover events occur at a much higher frequency.[33] Obviously, more backcrosses will increase the chance of a recombination event occurring closer to the causal mutation, thus increasing mapping resolution.



*100% if mutation resides on chromosome X
#If mutation resides on chromosome X, a 2-step backcross will only account for one backcross.Thus a 1-step back-cross strategy is recommended ≥6x

Figure 3. Flowchart of EMS-based simultaneous mapping and mutant identification using deep sequencing. This strategy for identifying a causal mutation traces the distinctive changes in nucleotides caused by EMS itself to map the location of the mutation. Several rounds of backcrossing remove EMS-induced nucleotide changes that are un-linked to the causal mutation, while leaving a "hot spot" of EMS damage surrounding the causal mutation (see ref. 12). Only a single recombinant is needed at the end of backcrossing to grow and extract DNA for deep sequencing. This is also advantageous when working with a mutant with a low penetrance or subtle phenotype. Deep sequencing data and analysis with available software[34] will reveal a genomic region linked to the causal mutation. Nucleotide mutations within this region will also be revealed and a list of candidate causal mutations that affect genes within this region will be assembled. The mutant can then be analyzed directly (since it has already been backcrossed) via several methods (e.g., complementation analysis and transgenic rescue) to determine which candidate mutation is causing the phenotype.

To further increase the number of recombination events that can be observed by deep sequencing, multiple recombinants can be analyzed together. Each recombinant obtained from a cross has undergone independent recombination events and, thus, analyzing several recombinants instead of one recombinant increases the resolution of mapping simply by increasing the number of chances that a recombination event has occurred more closely to the causal mutation. This is why SNP-based mapping protocols pool genomic DNA from multiple recombinant F_2 progeny before deep sequencing (see above, and **Fig. 1A**). In fact, this same bulk segregant approach can also be performed when using EMS-based mapping (**Fig. 1C**), as has recently been shown in both plants (rice) and *C. elegans*.[19,34] In *C. elegans*, it was found that this approach resulted in a mapping resolution (~1 Mb) very similar to that of SNP-based mapping.[34] In this example, linkage of background nucleotide variants present in the strain as a consequence of genetic drift were also followed to complement that of mutagen-induced nucleotide variants (variant density mapping).[34]

Thus, at the last round of backcrossing, which might be reduced to as little as one round if desired, multiple recombinant F_2 progeny can be selected, grown and pooled for DNA extraction (**Fig. 4**), much as in the manner as that performed for SNP-based mapping.[21] An added benefit of this approach is that at least one round of backcrossing has already been performed prior to mapping and sequencing, which will save time later, when more backcrosses are required prior to mutant characterization.

Advantages and Disadvantages of Each Strategy

Below we compare features of the SNP-based and EMS-based methods for simultaneously mapping and sequencing mutations, which are summarized in **Table 1** and can be used as a basis when deciding which strategy to implement. **Table 2** summarizes the considerations and choices along the way.

Selecting multiple vs. a single recombinant F_2. One scenario where it can be difficult to pool multiple recombinant progeny is when the mutant has a subtle phenotype, or has a very low penetrance or a combination of both. In our experience, with a mutant of around 5% penetrance and a subtle single-cell level phenotype, we found that it was very convenient that the isolation of only one recombinant was sufficient for mapping using the EMS-based approach (**Fig. 3**). This particular mutant was

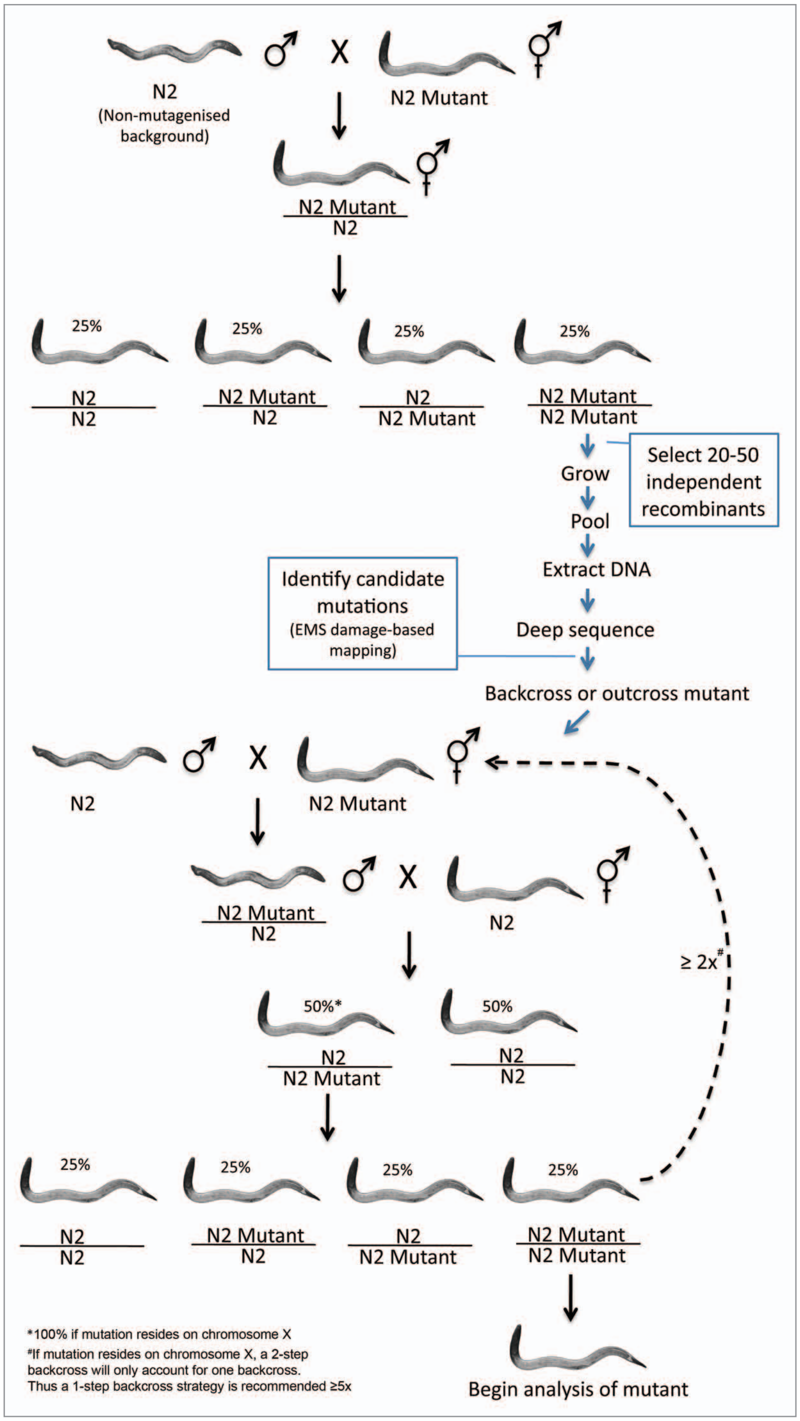

N2
(Non-mutagenised background)
♂
X
N2 Mutant
⚥
N2 Mutant / N2
25%
N2 / N2
N2 Mutant / N2
N2 / N2 Mutant
N2 Mutant / N2 Mutant
Select 20-50 independent recombinants
Grow
Pool
Extract DNA
Deep sequence
Identify candidate mutations (EMS damage-based mapping)
Backcross or outcross mutant
N2
N2 Mutant
N2 Mutant / N2
N2
50%*
N2 / N2 Mutant
50%
N2 / N2
≥ 2x#
Begin analysis of mutant
*100% if mutation resides on chromosome X
#If mutation resides on chromosome X, a 2-step backcross will only account for one backcross. Thus a 1-step backcross strategy is recommended ≥5x

Figure 4. Flowchart of a bulk segregation approach to EMS-based mapping using deep sequencing. Only a single backcross would be necessary to perform EMS-based mapping if multiple recombinant F_2 mutant progeny were singled, grown and pooled for deep sequencing. This approach has the added advantage of saving time later by reducing the number of backcrosses needed before mutant analysis as well as avoiding a CB4856 cross, which may preclude some mutant phenotypes. More than one backcross can be performed prior to selection of 20–50 recombinants, which would further increase mapping accuracy. Again, deep sequencing data and analysis with available software[34] will reveal a genomic region linked to the causal mutation. Nucleotide mutations within this region will also be revealed and a list of candidate causal mutations that affect genes within this region will be assembled. The mutant is then backcrossed, although a less number of times since it has already undergone a round of backcrossing, and then tested via several methods (e.g., complementation analysis and transgenic rescue) to determine which candidate mutation is causing the phenotype.

mapped to a section of chromosome V, where a candidate mutation within was confirmed to be causal (SZ and SJ, data not shown). This is a distinct advantage of EMS-based mapping when compared with SNP-based mapping and may also be useful when examining behavioral phenotypes. In such circumstances, recombinant progeny may need to be selected through complex analyses of behavior, which may need to be preceded by amplification of animal numbers. This is a challenging task that is further complicated by mutations with incomplete penetrance. Thus, proceeding through backcrossing, which is necessary anyway, and selecting a single recombinant at each step (**Fig. 3**) may be a more convenient strategy to map and sequence such behavioral mutants.

The same can also be said for mapping mutants from modifier genetic screens, where a second mutation (enhancer or suppressor) needs to be maintained together with the primary mutation. Indeed, a useful strategy would be to use the strain containing the primary mutation for backcrossing, thus ensuring that this mutation is maintained as homozygous throughout. This method would simplify identification of F_2 recombinants containing both mutations, which can be subsequently used for either single recombinant or bulk segregant EMS-based mapping (**Figs. 3 and 4**). This strategy cannot be achieved when outcrossing to CB4856 to create a mapping population.

Complications arising from background nucleotide variants. Another key difference between the two methods is that SNP-based mapping requires the introduction of variants (CB4856 SNPs) into the background, whereas EMS-based mapping requires the removal of variants (caused by EMS) from the background. For some phenotypes, it is possible that altering the genetic background via the introduction of CB4856 polymorphisms may modify the phenotype. As such, it would not be possible to create a mapping population by crossing to CB4856.

Comparisons between mutant sequences. Discerning between false nucleotide variants that may have been called as a result from comparison to the N2 reference genome (which contains rare sequencing errors, ~1/100,000 nucleotides[13-15]), background variants specific to the strain used in the mutagenesis screen and nucleotide changes induced by EMS is a key aspect of successfully identifying bona fide candidate mutations. Even within a small genomic region mapped to a few million base pairs, there are still many background variants that hamper the selection of candidate mutations. Sequencing and directly

Table 1. Advantages and disadvantages of SNP-based and EMS-based mapping strategies

Method		Advantage	Disadvantage
SNP-based approach		• Out/backcrossing not necessary for mapping mutation • CB4856 strain readily available for mapping population • Mapping population can be used for rough mapping prior to deep sequencing • Crossing step quicker than multiple rounds of out/backcrossing • Introduced CB4856 polymorphisms may modify some phenotypes (potentially interesting)	• Need to create a mapping population (20–50 F_2) • Still requires out/backcrossing after candidate mutations identified • Need to sequence background strain separately as comparison directly to other mutant sequences hindered by recombinant CB4856 genome • Introduced CB4856 polymorphisms may preclude some phenotypes
EMS-based approach	Non-bulk segregant approach	• No need to create a mapping population • Only one animal needs to be recovered for each backcross step and for sequencing • Out/backcrossing is inherent and therefore does not need to be performed later • Comparison between multiple mutant isogenic sequences possible • Crossing to the same strain won't preclude phenotype • Useful for cloning modifier mutations	• Necessary out/backcrossing takes a longer (horizontal) time • EMS-induced nucleotide changes less frequent than polymorphisms (although not limiting factor to mapping resolution) • Mapping resolution more variable due to lower number of recombination events • Possibility of low EMS mutation density surrounding causal mutation, which may reduce mapping resolution
	Bulk-segregant approach	• Mapping resolution greater than non-bulk segregant approach • Crossing step quicker than multiple rounds of out/backcrossing • Reduced number of out/backcrosses to be performed later • Comparison between multiple mutant isogenic sequences possible • Crossing to the same strain won't preclude phenotype • Useful for cloning modifier mutations	• Need to create a mapping population (20–50 F_2) • EMS-induced nucleotide changes less frequent than polymorphisms (although not limiting factor to mapping resolution) • Still requires more out/backcrossing after candidate mutations identified • Possibility of low EMS mutation density surrounding causal mutation, which may reduce mapping resolution

Table 2. Step by step practical considerations and choices when implementing deep sequencing strategies for mapping and identifying mutations

Choose a mapping strategy	Impinges on whether additional genetic tests will be necessary (see selecting multiple vs. a single recombinant F2, complications arising from background nucleotide variants and time involved)
Choose how many alleles will be sequenced in parallel	(see comparisons between mutant sequences and ability to prioritize mutants to analyze)
Determine the depth of sequencing	impinges on the accurate calling of variants (see limitations to deep sequencing for mutant identification)
Choose an analysis platform	(see analyzing deep sequencing data). The parameters and filtering stringency used will impact on the ability to make accurate variant calls
Identify candidates for the causal mutation	- start looking for mutations altering the coding sequence or splice donor/acceptor sites - filter out possible sequencing errors or variants from the N2 reference sequence - look at all variants in the mapped region (i.e., point mutations vs. indels, EMS-canonical vs. non-canonical changes) (see analyzing deep sequencing data and limitations to deep sequencing for mutant identification)

comparing multiple mutants of the same genetic background in parallel, or in independent sequencing experiments, makes identifying candidate mutations a much simpler task as shared variants between the sequenced strains can be discounted. This is relatively straightforward when performing EMS-based mapping, as mutants are moved toward an isogenic state through multiple rounds of backcrossing. Of course, in order not to discount a true causal mutation that happens to present itself as the exact same DNA lesion in two separate mutants (a rare occurrence), it is important to confirm that each variant being discounted occurs also in separate mutants that map to different regions of the genome.

However, for SNP-based mapping, where sequence comparison may be hindered between mutants of the same screen due to the introduction of stretches of CB4856 genomic DNA, the premutagenized strain may be sequenced separately for comparison. In addition, subtraction of already known and published nucleotide variants in *C. elegans* will help to facilitate the identification

Table 3. Latest technology and costs of deep sequencing (adapted from Hayden 2013[43])

Technology	Cost of machine	Read length	Cost per Mb
Ion Torrent Proton	$224,000 USD	200 bp	1–9 cents
Illumina HiSeq	$690,000 USD	300 bp	4–5 cents
Illumina MiSeq	$125,000 USD	500 bp	14–70 cents
Ion Torrent PGM	$49,000 USD	400 bp	60 cents–$5 USD
PacBio RS	$695,000 USD	4,575 bp	$2–17 USD

While multiplexing and increased reads capacity per chip have enabled the sequencing costs to lower, the costs associated with the preparation of the sequencing library have remained stable. The number of times the genome will be sequenced, called the sequencing depth or coverage, dictates the total cost. Coverage can be calculated as the read length × the number of reads × 1 or 2 depending on whether single or paired ends are sequenced, over the haploid genome length. Because reads are not distributed evenly over the genome, many bases will be covered by more reads than the average aimed for while others will be covered by fewer reads. In addition, certain regions are harder to sequence. In our experience, a coverage of 20X–30X is sufficient for accurate mutation identification.

of putative phenotype-causing mutation [for example, see file **Fig. S1**,[12] **Table S1**,[15] WormBase (www.wormbase.org)[35]].

Time involved. One drawback of backcrossing or outcrossing multiple times compared with a single cross with CB4856 to create a mapping population is the horizontal time required. Four or more crosses to the original un-mutagenized strain can take a month or more, although as said before, it is possible to make fewer backcrosses–for instance, just one backcross–and select and pool multiple F_2 recombinants for DNA sequencing (**Fig. 4**). In either case, any mapping strategy will not avoid the need to eventually backcross or outcross mutagenized animals in order to remove background mutations that could confound the interpretation of the impact of the primary mutation. In fact, it would be advisable under any circumstances, before investing in any directed efforts to sequence, to first backcross or outcross a mutant one or several times to determine if the mutation behaves in a Mendelian fashion, is a single locus, is linked to chromosome X and is recessive or dominant.

Ability to prioritize mutants to analyze. If one wishes to first prioritize mutants from a screen to be sequenced based on their likelihood of affecting distinct loci, then generating a mapping population with CB4856 may provide an advantage. Although somewhat defeating the purpose of simultaneously mapping and sequencing a mutation, rough mapping (using traditional techniques) of the mapping populations prior to sequencing might hint that multiple mutants affect the same gene and, thus, only one of these could be sequenced. Caution should be used with such an approach since rough mapping is indeed a very coarse method to localize a mutation to an approximate chromosomal region containing many genes. Alternatively, complementation analysis of each mutant from a screen can be employed without the need for a mapping population, and is a much more accurate method to assess whether multiple mutants of the same screen affect the same gene. However, in light of the spiraling downward costs of deep sequencing (see next section), it may not be worthwhile at all to choose which mutants to sequence, but rather to sequence them all, saving both time, reagents and labor. Indeed, sequencing multiple mutant alleles of the same gene would be very useful for immediate identification of candidate mutations, and can be very informative for the function of the gene, advantages that outweigh redundancy concerns.

Cost of Whole Genome Sequencing

Currently, a mutant *C. elegans* genome can be sequenced to 20X fold coverage for less than $500 USD. Reducing prices can be attributed to advances in deep sequencing technology that allow a greater density of DNA clusters on the surface of each flow cell lane of a sequencing chip, as well as longer read lengths and analysis of multiple independent genomes on a single flow lane, multiplexing.[36] Over the past decade, the plummeting costs of DNA sequencing have exceeded Moore's law, which describes a long-term trend in the computer hardware industry of a doubling of power every 2 y.[37] **Table 3** provides a summary of the latest sequencing platforms and their associated costs.

Analyzing Deep Sequencing Data

Analyzing deep sequencing results from *C. elegans* was aided by the development of programs such as MAQgene,[38] which uses the assembly building software MAQ.[39] Although once useful, MAQ is now an outdated aligner and installation of MAQgene required specialized bioinformatic expertise. More recently, a cloud-based pipeline called CloudMap was developed, which is entirely Internet-based and therefore does not require any software installation.[34] This latest development allows users to choose from several aligners including the Burrows-Wheeler Aligner (BWA)[40] and Bowtie.[41] With CloudMap, a specific map position and list of candidate variants from raw sequencing data from all major NGS platforms (Illumina, ABI, 454) can be obtained using either of the different mapping strategies discussed in this review (EMS-based and SNP-based mapping). Alternatively, SAMtools (Sequence Alignment/Map)[42] may be used for alignment and nucleotide variant identification but does not incorporate the automated workflow for SNP- or EMS-based mapping outputs that CloudMap does. Readers are referred to the primary papers of each of these alignment programs for optimization of parameters such as mapping quality and base quality thresholds as well as read depth filters to alter the stringency of nucleotide variant lists generated. Calling of bona fide deletions (see below) is particularly sensitive to the settings used, as deleted sections of DNA can be interpreted as uncovered regions of sequence.

Once a mapping interval has been determined (a graphical output is presented by CloudMap), it is then up to the user to discriminate those mutations in the mapping region that are the strongest candidates to examine in priority. It is likely that the causal mutation will affect the coding sequence of a given gene (a description of which will be in the output list), generating a missense, nonsense or a splicing defect. However, it is entirely possible that mutations in regulatory regions upstream, downstream or within introns may also cause the phenotype and, therefore, should not be rashly discounted. It is also important to take into account that EMS does not only induce canonical G to A nucleotide transitions (C to T on the opposite strand), but also other DNA lesions including deletions. Although occurring at a far lesser frequency than point mutations (which are not only restricted to G > A changes[14]), these EMS-induced deletions can be quite large (e.g., 1,888 bp in *C. elegans*).[11] Thus, when a mapping region has been determined, a thorough examination of all nucleotide transitions and tranversions, as well as an investigation of potential deletions, which may be called as "uncovered regions" (depending on the aligning software and settings used), should take place in this region. True deletions or insertions can be confirmed as such by PCR and Sanger sequencing, and if coverage of a particular nucleotide variant is extremely low (e.g., < 3-fold) the same can be done for point mutations. Of course, if the same "uncovered region" occurs in an independent strain or in an independent mutant of the same background, it likely reflects either a difficult region to sequence or align (e.g., repetitive sequences) or a deletion initially present in the starting strain and is therefore not causal. Confirmation of a causal mutation can be achieved via phenocopy (using RNAi or available mutants), complementation analysis and transgenic rescue.

Limitations to Deep Sequencing for Mutant Identification

One potential limitation to the successful cloning of a mutant using deep sequencing arises if the causal DNA lesion is not covered by the sequences obtained. However, the chances of this occurring are quite slim if enough depth in sequence is obtained. Flibotte et al. (2010) calculated that the overall sensitivity to detect point mutations was 89% when sequencing to a depth of 20X (the average number of times each nucleotide was sequenced across the genome). This was improved to 95% when focusing only on exons, in which most causal mutations are likely to occur in.[14] Another potential problem relates to the potential of false negative or false positive miscalls of nucleotide variants, especially of poorly covered or difficult to sequence genomic regions. The simultaneous mapping and deep sequencing strategies discussed here are relatively robust to individual variant miscalls, as they each rely on the combined data of many variant calls (or lack thereof). However, when examining in detail the mapping region to identify candidate mutations, miscalling of nucleotide variants may prove crucial in identifying the true causal mutation.

Phenotypes arising from gene duplication may also present a problem when trying to identify a genetic cause. This type of DNA lesion would be quite difficult to detect with deep sequencing. The number of reads obtained for a given section of genome is very variable and, as such, the copy number of a gene would be very difficult to ascertain. Although, we have observed that integrated, multi-copy transgenes present as a greater number of sequence reads than neighboring genomic regions, a single duplication event would be almost impossible to detect in this manner. It is possible however, that the individual sequence reads that cover each end of a duplicated gene will overlap with an entirely different region of genome than that of the genes' native position. This would hint at a possible duplication event as well as the location of the duplication within the genome.

Finally, it may be that the causal mutation lies within a misannotated region (cDNA or exon not annotated) of the *C. elegans* genome. In such cases, we have found it very useful to verify the degree of conservation across different nematode genomes around putative mutations when these are not localized to a known ORF [for example, using the UCSC genome browser (genome.ucsc.edu/)].

Conclusion

Maturation of next generation sequencing technology has enabled affordable and rapid whole genome sequencing. However, when using this technology for forward genetic screens, it is important to consider that identification of total nucleotide variants alone will not, in the majority of cases, allow a small enough list of candidate mutations to be practically tested. Instead, many background variants across the genome–including those that affect coding sequences–will be revealed, potentially confounding the identification of the true causal mutation. For this reason, several methods have been developed to simultaneously deep sequence and map the causal mutation to a region of the genome, aiding the identification of the causal mutation. These strategies have dramatically shrunk the time needed to identify a causal mutation from a mutagenesis screen. As such, the steps involved in identifying a causal mutation(s) no longer limit forward genetic mutagenesis screens. Depending upon the particular screen to be performed and the characteristics and number of the resulting mutants, one or a combination of the alternative mapping strategies discussed here will be best suited to quickly and effectively identify the causal mutation and gene behind the phenotype.

Disclosure of Potential Conflicts of Interest

No potential conflicts of interest were disclosed.

Acknowledgments

We acknowledge the helpful comments of members of the Jarriault Lab as well as Oliver Hobert. This work was supported by a Fondation ARC (Association pour la Recherche sur le Cancer) Postdoctoral Fellowship (to S.Z.) and grants from the Centre National de la Recherche Scientifique-ATIP (Action Thématique Incitative sur Programme), the Fondation pour la Recherche Médicale (FRM-DRC20091217181), and European Molecular Biology Organization (EMBO) Young Investigator Prize (to S.J.). S.J. is an investigator of the Centre National de la Recherche Scientifique.

References

1. Jorgensen EM, Mango SE. The art and design of genetic screens: caenorhabditis elegans. Nat Rev Genet 2002; 3:356-69; PMID:11988761; http://dx.doi.org/10.1038/nrg794
2. Kamath RS, Ahringer J. Genome-wide RNAi screening in Caenorhabditis elegans. Methods 2003; 30:313-21; PMID:12828945; http://dx.doi.org/10.1016/S1046-2023(03)00050-1
3. Davis MW, Hammarlund M, Harrach T, Hullett P, Olsen S, Jorgensen EM. Rapid single nucleotide polymorphism mapping in C. elegans. BMC Genomics 2005; 6:118; PMID:16156901; http://dx.doi.org/10.1186/1471-2164-6-118
4. Wicks SR, Yeh RT, Gish WR, Waterston RH, Plasterk RH. Rapid gene mapping in Caenorhabditis elegans using a high density polymorphism map. Nat Genet 2001; 28:160-4; PMID:11381264; http://dx.doi.org/10.1038/88878
5. Williams BD, Schrank B, Huynh C, Shownkeen R, Waterston RH. A genetic mapping system in Caenorhabditis elegans based on polymorphic sequence-tagged sites. Genetics 1992; 131:609-24; PMID:1321065
6. Shendure J, Ji H. Next-generation DNA sequencing. Nat Biotechnol 2008; 26:1135-45; PMID:18846087; http://dx.doi.org/10.1038/nbt1486
7. Meyerson M, Gabriel S, Getz G. Advances in understanding cancer genomes through second-generation sequencing. Nat Rev Genet 2010; 11:685-96; PMID:20847746; http://dx.doi.org/10.1038/nrg2841
8. Veltman JA, Brunner HG. De novo mutations in human genetic disease. Nat Rev Genet 2012; 13:565-75; PMID:22805709; http://dx.doi.org/10.1038/nrg3241
9. Harrison PW, Wright AE, Mank JE. The evolution of gene expression and the transcriptome-phenotype relationship. Semin Cell Dev Biol 2012; 23:222-9; PMID:22210502; http://dx.doi.org/10.1016/j.semcdb.2011.12.004
10. Sarin S, Prabhu S, O'Meara MM, Pe'er I, Hobert O. Caenorhabditis elegans mutant allele identification by whole-genome sequencing. Nat Methods 2008; 5:865-7; PMID:18677319; http://dx.doi.org/10.1038/nmeth.1249
11. Sarin S, Bertrand V, Bigelow H, Boyanov A, Doitsidou M, Poole R, et al. Analysis of Multiple EMS-mutagenized Caenorhabditis elegans Strains by Whole Genome Sequencing. Genetics 2010; 185:417-30; PMID:20439776; http://dx.doi.org/10.1534/genetics.110.116319
12. Zuryn S, Le Gras S, Jamet K, Jarriault S. A strategy for direct mapping and identification of mutations by whole-genome sequencing. Genetics 2010; 186:427-30; PMID:20610404; http://dx.doi.org/10.1534/genetics.110.119230
13. Hillier LW, Marth GT, Quinlan AR, Dooling D, Fewell G, Barnett D, et al. Whole-genome sequencing and variant discovery in C. elegans. Nat Methods 2008; 5:183-8; PMID:18204455; http://dx.doi.org/10.1038/nmeth.1179
14. Flibotte S, Edgley ML, Chaudhry I, Taylor J, Neil SE, Rogula A, et al. Whole-genome profiling of mutagenesis in Caenorhabditis elegans. Genetics 2010; 185:431-41; PMID:20439774; http://dx.doi.org/10.1534/genetics.110.116616
15. Weber KP, De S, Kozarewa I, Turner DJ, Babu MM, de Bono M. Whole genome sequencing highlights genetic changes associated with laboratory domestication of C. elegans. PLoS One 2010; 5:e13922; PMID:21085631; http://dx.doi.org/10.1371/journal.pone.0013922
16. Denver DR, Morris K, Lynch M, Thomas WK. High mutation rate and predominance of insertions in the Caenorhabditis elegans nuclear genome. Nature 2004; 430:679-82; PMID:15295601; http://dx.doi.org/10.1038/nature02697
17. Obholzer N, Swinburne IA, Schwab E, Nechiporuk AV, Nicolson T, Megason SG. Rapid positional cloning of zebrafish mutations by linkage and homozygosity mapping using whole-genome sequencing. Development 2012; 139:4280-90; PMID:23052906; http://dx.doi.org/10.1242/dev.083931
18. Schneeberger K, Ossowski S, Lanz C, Juul T, Petersen AH, Nielsen KL, et al. SHOREmap: simultaneous mapping and mutation identification by deep sequencing. Nat Methods 2009; 6:550-1; PMID:19644454; http://dx.doi.org/10.1038/nmeth0809-550
19. Abe A, Kosugi S, Yoshida K, Natsume S, Takagi H, Kanzaki H, et al. Genome sequencing reveals agronomically important loci in rice using MutMap. Nat Biotechnol 2012; 30:174-8; PMID:22267009; http://dx.doi.org/10.1038/nbt.2095
20. Bull KR, Rimmer AJ, Siggs OM, Miosge LA, Roots CM, Enders A, et al. Unlocking the bottleneck in forward genetics using whole-genome sequencing and identity by descent to isolate causative mutations. PLoS Genet 2013; 9:e1003219; PMID:23382690; http://dx.doi.org/10.1371/journal.pgen.1003219
21. Doitsidou M, Poole RJ, Sarin S, Bigelow H, Hobert O. C. elegans mutant identification with a one-step whole-genome-sequencing and SNP mapping strategy. PLoS One 2010; 5:e15435; PMID:21079745; http://dx.doi.org/10.1371/journal.pone.0015435
22. Lister R, Gregory BD, Ecker JR. Next is now: new technologies for sequencing of genomes, transcriptomes, and beyond. Curr Opin Plant Biol 2009; 12:107-18; PMID:19157957; http://dx.doi.org/10.1016/j.pbi.2008.11.004
23. Hodgkin J, Doniach T. Natural variation and copulatory plug formation in Caenorhabditis elegans. Genetics 1997; 146:149-64; PMID:9136008
24. Lewis ZA, Shiver AL, Stiffler N, Miller MR, Johnson EA, Selker EU. High-density detection of restriction-site-associated DNA markers for rapid mapping of mutated loci in Neurospora. Genetics 2007; 177:1163-71; PMID:17660537; http://dx.doi.org/10.1534/genetics.107.078147
25. Miller MR, Dunham JP, Amores A, Cresko WA, Johnson EA. Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers. Genome Res 2007; 17:240-8; PMID:17189378; http://dx.doi.org/10.1101/gr.5681207
26. O'Rourke SM, Yochem J, Connolly AA, Price MH, Carter L, Lowry JB, et al. Rapid mapping and identification of mutations in Caenorhabditis elegans by restriction site-associated DNA mapping and genomic interval pull-down sequencing. Genetics 2011; 189:767-78; PMID:21900274; http://dx.doi.org/10.1534/genetics.111.134031
27. Jarriault S, Schwab Y, Greenwald I. A Caenorhabditis elegans model for epithelial-neuronal transdifferentiation. Proc Natl Acad Sci USA 2008; 105:3790-5; PMID:18308937; http://dx.doi.org/10.1073/pnas.0712159105
28. Richard JP, Zuryn S, Fischer N, Pavet V, Vaucamps N, Jarriault S. Direct in vivo cellular reprogramming involves transition through discrete, non-pluripotent steps. Development 2011; 138:1483-92; PMID:21389048; http://dx.doi.org/10.1242/dev.063115
29. Kagias K, Ahier A, Fischer N, Jarriault S. Members of the NODE (Nanog and Oct4-associated deacetylase) complex and SOX-2 promote the initiation of a natural cellular reprogramming event in vivo. Proc Natl Acad Sci USA 2012; 109:6596-601; PMID:22493276; http://dx.doi.org/10.1073/pnas.1117031109
30. Coulondre C, Miller JH. Genetic studies of the lac repressor. III. Additional correlation of mutational sites with specific amino acid residues. J Mol Biol 1977; 117:525-67; PMID:609095; http://dx.doi.org/10.1016/0022-2836(77)90056-0
31. Bautz E, Freese E. On the Mutagenic Effect of Alkylating Agents. Proc Natl Acad Sci USA 1960; 46:1585-94; PMID:16590785; http://dx.doi.org/10.1073/pnas.46.12.1585
32. Meneely PM, Farago AF, Kauffman TM. Crossover distribution and high interference for both the X chromosome and an autosome during oogenesis and spermatogenesis in Caenorhabditis elegans. Genetics 2002; 162:1169-77; PMID:12454064
33. Martinez-Perez E, Colaiácovo MP. Distribution of meiotic recombination events: talking to your neighbors. Curr Opin Genet Dev 2009; 19:105-12; PMID:19328674; http://dx.doi.org/10.1016/j.gde.2009.02.005
34. Minevich G, Park DS, Blankenberg D, Poole RJ, Hobert O. CloudMap: a cloud-based pipeline for analysis of mutant genome sequences. Genetics 2012; 192:1249-69; PMID:23051646; http://dx.doi.org/10.1534/genetics.112.144204
35. Yook K, Harris TW, Bieri T, Cabunoc A, Chan J, Chen WJ, et al. WormBase 2012: more genomes, more data, new website. Nucleic Acids Res 2012; 40(Database issue):D735-41; PMID:22067452; http://dx.doi.org/10.1093/nar/gkr954
36. Cronn R, Liston A, Parks M, Gernandt DS, Shen R, Mockler T. Multiplex sequencing of plant chloroplast genomes using Solexa sequencing-by-synthesis technology. Nucleic Acids Res 2008; 36:e122; PMID:18753151; http://dx.doi.org/10.1093/nar/gkn502
37. Dewitt ND, Yaffe MP, Trounson A. Building stem-cell genomics in California and beyond. Nat Biotechnol 2012; 30:20-5; PMID:22231086; http://dx.doi.org/10.1038/nbt.2086
38. Bigelow H, Doitsidou M, Sarin S, Hobert O. MAQGene: software to facilitate C. elegans mutant genome sequence analysis. Nat Methods 2009; 6:549; PMID:19620971; http://dx.doi.org/10.1038/nmeth.f.260
39. Li H, Ruan J, Durbin R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res 2008; 18:1851-8; PMID:18714091; http://dx.doi.org/10.1101/gr.078212.108
40. Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics 2010; 26:589-95; PMID:20080505; http://dx.doi.org/10.1093/bioinformatics/btp698
41. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 2009; 10:R25; PMID:19261174; http://dx.doi.org/10.1186/gb-2009-10-3-r25
42. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al.; 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. Bioinformatics 2009; 25:2078-9; PMID:19505943; http://dx.doi.org/10.1093/bioinformatics/btp352
43. Hayden EC. Gene sequencing leaves the laboratory. Nature 2013; 494:290-1; PMID:23426300; http://dx.doi.org/10.1038/494290a