Population Sequencing for Studying Natural and Artificial Variation in *C. elegans*

by

Brad T. Moore

Graduate Program in Computational Biology and Bioinformatics
Duke University

Date: ________________________
Approved:

____
L. Ryan Baugh, Supervisor

____
Timothy E. Reddy

____
Sayan Mukherjee

____
Paul M. Magwene

____
Philip N. Benfey

Dissertation submitted in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in the Graduate Program in Computational Biology and
Bioinformatics
in the Graduate School of Duke University
2017

ABSTRACT

Population Sequencing for Studying Natural and Artificial Variation in *C. elegans*

by

Brad T. Moore

Graduate Program in Computational Biology and Bioinformatics
Duke University

Date: ____________________
Approved:

____
L. Ryan Baugh, Supervisor

____
Timothy E. Reddy

____
Sayan Mukherjee

____
Paul M. Magwene

____
Philip N. Benfey

An abstract of a dissertation submitted in partial fulfillment of the requirements for
the degree of Doctor of Philosophy in the Graduate Program in Computational
Biology and Bioinformatics
in the Graduate School of Duke University
2017

Abstract

The advent of high coverage and low cost sequencing technologies has allowed for newer and more powerful approaches in molecular and population genetics. Transposon sequencing, where genome-saturated mutant populations allele frequencies are measured before and after selection, functionally characterizes each and every gene in the genome in a single experiment. The approach has been successfully applied to a variety of phenotypes in a variety of unicellular systems: growth and motility in *E. coli*, synthetic genetic interactions in yeast, and *in vitro* pathogen-resistance in mammalian cell lines. However, transposon insertion typically produces null alleles, which can be valuable to identify gene function, but evolutionary insight relies on identification of naturally occurring polymorphisms affecting the trait of interest. Genome-wide association studies (GWAS) can be used to study the effect of natural genetic variation on a trait, but they grow prohibitively expensive if the number of individuals to genotype and phenotype becomes large.

Here I describe the application of transposon sequencing and pooled sequencing GWAS in the whole metazoan model, *Caenorhabditis elegans.* Transposon sequencing has not been previously implemented in an animal model. I have sequenced a control library using our method, *C. elegans* transposon sequencing (CeTnSeq). We have constructed a new Mos1 transposon mutator strain that is more convenient to use than the existing strain and allows for extra-chromosomal insertions to be degraded by restriction digest. My preliminary results show that our method is qualitatively effective at identifying transposon insertion sites, but suffers from PCR duplication error. I propose to optimize the number of PCR cycles in the library and to include unique molecular identifiers (UMI) in the library adaptor. I also show that the restriction digest is effective at removing extra-chromosomal array insertions from the library.

I constructed simulation models to help design optimal Ce-TnSeq experiments with respect to statistical power for a proposed starvation survival assay. I considered many parameters affecting the design, including: culture size, number of generations, expected effect size, sequencing coverage, and sample size. I show that the number of homozygous mutant animals in the screen is a critical factor in the design of experiments. I also saw diminishing returns with respect to increasing sample size and sequencing depth. These simulations will be invaluable in designing future Ce-TnSeq experiments and identifying critical aspects of the protocol to optimize.

We performed pooled sequencing (using restriction-site associated DNA sequencing) on a population of 95 wild isolates subjected to starvation. I identified strains that were resistant and sensitive to starvation, and we verified these results using traditional methods. We used our population sequencing data to perform an association study of starvation survival across the 95 strains, and identified two statistically significant quantitative trait loci.

To my wife Megan and son Jacob- let the good times roll.

Contents

List of Tables

List of Figures

Acknowledgements

I would like to acknowledge the following funding sources: National Science Foundation (NSF T32 (Hartemink/Harer)), National Institutes of Health (NIH R01 (Baugh), R21 (Baugh)), Duke Translational Research Institute (HiSeq voucher, RADseq).

I would also like to acknowledge contributors to this work, specifically: Beth Bowman for prototype experiments in Ce-TnSeq, Becky Kaplan for creating standard curves for Ce-TnSeq, Jim Jordan for running the first Ce-TnSeq assays, Rojin Chitrakar for running the first Ce-TnSeq control assay and performing the injections for LRB311, Nicolas Devos for providing Illumina sequencing support, Tim Reddy for helping troubleshoot the first Ce-TnSeq sequencing runs, Paul Magwene for quantitative support on the Ce-TnSeq modeling, Sayan Mukherjee for statistical support for the RADseq project, and Anthony Hung for significant contributions to the RADseq project (maintenance of the 97 strains, pooling of the strains, and the validation experiment).

I would also like to thank Ryan Baugh, who was my advisor while at Duke.

1

Introduction

1.1 C. elegans as a Model Organism

Caenorhabditis elegans is a nematode model organism. It is typically a hermaphrodite (males occur at rate less that 0.001% in the laboratory strain) and reproduces primarily by self-crossing. Its diploid genome is 100 mb in size and has approximately 20K protein-coding genes, 38% of which have human orthologs (Shaye and Greenwald (2011)). It grows quickly, only taking 60 hours to develop from a laid egg to a reproductive adult. It is amenable to genetic transformation: plasmids injected into the germline of a young adult are non-homology repaired into high-copy number artificial chromosomes called extra-chromosomal arrays. The usual genetic tools exist in the worm, including: RNA interference (by feeding with dsRNA producing bacteria), CRISPR, fluorescent reporters, etc (Ahringer (2006)). The worm is maintained on agar plates seeded with *E. coli*, or can be grown at large scale (millions of worms) in liquid culture. The worm has a well-conserved cell lineage, with L1 larva having exactly 558 cells and adult worms having approximately 1000 somatic cells plus a variable number of germline cells (Sulston et al. (1983)).

During normal development, the worm passes through four larval stages (L1-L4, molting in-between). It has an alternative larval stage called dauer, which it may enter between L1 and L2 larval stages due to low nutrition or overcrowding (dauer pheromone). If hatched in the absence of food, the worm arrests development in the first larval stage (called L1 arrest or L1 diapause). These alternative developmental plans are important in the study of aging and metabolism, and orthologs of the human insulin-like receptor (DAF-2 in the worm) and forkhead transcription factor FOXO (DAF-16) are key regulators of dauer and L1 arrest (Baugh et al. (2009); Kenyon et al. (1993)).

1.2 Insertional Mutagenesis

Forward genetic screens have been a cornerstone of genetic research in model organisms (St Johnston (2002); Page and Grossniklaus (2002); Jorgensen and Mango (2002)). They allow for the genes controlling a particular phenotype to be discovered without prior knowledge. Insertional mutagens, known elements of DNA which are randomly integrated into a genome, allow for rapid and large-scale genotyping of mutants. The known DNA sequence of insertional mutagens can be leveraged to selectively sequence the regions adjacent to a mutation and identify its location. Several types of insertional mutagens exist, such as transposons and transfer DNA (T-DNA; Boulin and Bessereau (2007), Desfeux et al. (2000)). Typically, when an insertional mutagen is integrated into an endogenous gene, it creates a loss-of-function mutation, though gain-of-function mutants have been created with enhancer-containing elements (Weigel et al. (2000)). Overall, the expected effect of an insertional mutant depends on the type of mutation (loss or gain-of-function) and the ploidy of the organism. In general, loss-of-function mutations are expected to be recessive as a functional copy of the gene on the sister chromosome may compensate, and gain-of-function mutations are expected to be dominant.

In haploid organisms, the design of an insertional mutant screen is similar to that of chemical mutagens and is affected mainly by the mutation rate, culture size, and labor of the screen (Shuman and Silvahy (2003)). In diploid organisms, however, a way to homozygous insertional mutants must be considered. In *Drosophila melanogaster*, heterozygous male mutants are backcrossed to virgin females containing a balancer (a selectable and recessively lethal marker) for the chromosome under study (St Johnston (2002)). The F2 animals heterozygous for the balancer and mutation are then inbred thereby producing F3 progeny that are either heterozygous for the balancer and mutation, or homozygous for the mutation. In *Arabidopsis thaliana*, heterozygous mutant F1 animals are self-crossed to produce homozygous F2 progeny (Page and Grossniklaus (2002)). The insertional mutagen (either transposon or T-DNA) contained a selectable marker to select for mutants.

In *C. elegans*, endogenous (Tc1, Tc3, Tc5; Martin et al. (2002)) and exogenous (Mos1; Boulin and Bessereau (2007)) *Tc1/mariner*-family transposons have been used for insertional mutagenesis. Unlike the selectable-marker containing insertional mutagens used in *Drosophila* and *Arabidopsis*, these transposons were used in their native form. Mutant F1 worms were self-crossed, similarly to *Arabidopsis*, in order to produce homozygous mutant F2 animals. However, without selection, homozygous wild-type animals would continue into the screen (Boulin and Bessereau (2007)).

1.3 Transposon Sequencing

Transposon sequencing (TnSeq, also known as insertion sequencing or INSeq) has been a recent development that has dramatically increased our ability to conduct genetic screens (van Opijnen et al. (2009); Goodman et al. (2012); Gallagher et al. (2011); Rinaldi et al. (2012); Bronner et al. (2016)). The general idea is to create a saturated insertional mutant population, subject it to some selection, and use sequencing to identify changes in mutant allele frequencies. In other cases, the lack of detectable mutation in genomic regions has been used to determine the essentiality of genes (van Opijnen et al. (2009)). The technique relies on the ability to create high-throughput sequencing libraries from *only* the sites of mutation, leveraging the known sequence of the insertional mutagen. The measure of allele frequency (derived from number of sequencing reads) is then a quantitative measure of the functional effect of each mutant in the screen. Unlike traditional forward genetic screens, where only a handful of mutant lines are recovered, transposon sequencing can provide a measure of gene function for each and every gene in the genome in a single experiment.

A key challenge of transposon sequencing is the creation of the sequencing library. The original TnSeq method (van Opijnen et al. (2009)) leveraged an MmeI (a class IIS restriction enzyme that cuts 20bp downstream of its target) cutsite found in the inverted terminal repeat (ITR) of the magellan6 mariner transposon. By performing an MmeI digest on genomic DNA containing magellan6 insertions they were able to ligate sequencing adaptors to only those insertion sites. Other methods (see Table 1.1) have used probe-mediated ligation (Langridge et al. (2009)), *in vitro* transcription and microarray hybridization (Sassetti et al. (2001)), and transposon-specific PCR with blunt adaptor ligation (Bronner et al. (2016)).

Transposon sequencing has been performed in prokaryotes (van Opijnen et al. (2009); Goodman et al. (2012); Langridge et al. (2009); Sassetti et al. (2001); Gallagher et al. (2011); Gawronski et al. (2009)) and single-celled eukaryotes (Bronner et al. (2016)). To date, it has not be performed in an animal model. Note, Rinaldi et al. (2012) performed pooled sequencing of *Schitsosoma mansoni* insertional mutants, however, their method only qualitatively characterized the insertion bias of the mutagen used. They did not perform a saturating screen nor did they use their pooled sequencing results to infer gene function.

The type of analysis used on transposon sequencing reads depends on the type of screen. For essential genes in a saturated screen, hidden Markov model (Solaimanpour et al. (2015)) or window-based (Dejesus et al. (2015)) methods have been used to detect essential regions of the genome that lack any insertion. For conditionally essential gene, the significant changes in allele frequency after selection have been detected using differential expression tools (e.g. *edgeR*; Dembek et al. (2015)) or hidden Markov models (Pritchard et al. (2014)).

	Citation	Organism	Library Type	Mutagenesis	Study
1	van Opijnen et al. (2009)	Streptococcus pneumoniae	TnSeq (MmeI type IIS digest and PCR), Illumina	mariner (magellan6)	Single gene essentiality
2	Goodman et al. (2012)	Bacteroides thetaiotaomicron	INSeq (MmeI type IIS digest, linear PCR, biotin), Illumina	mariner (HimarI-derived)	Different growth conditions
3	Gawronski et al. (2009)	Haemophilus influenzae	HITS (Illumina adaptor ligation, junction PCR, biotin)	mariner (HimarI-derived)	Selecting for survivability in host
4	Langridge et al. (2009)	Salmonella typhi	TraDIS (Illumina ligation with single PCR)	Tn5	Growth in presence of bile
5	Sassetti et al. (2001)	Mycobacterium bovis (tuberculosis)	TraSH (transduced with phage, T7 to make RNA, hybridized to microarray)	mariner (HimarI-derived)	Different growth conditions
6	Bronner et al. (2016)	Plasmodium falciparum (malaria)	Qiseq (Illumina, splinkerette, nested PCR)	piggyBac	Fitness
7	Gallagher et al. (2011)	Pseudomonas aeruginosa	Tnseq-circle (Illumina, adaptor ligation, ligation-based circularization)	T8 (Tn5-derived)	Essentiality and drug resistance
8	Rinaldi et al. (2012)*	Schitsosoma mansoni (flatworm)	Illumina, ligated adaptors and nested PCR	MLV retroviral transduction	Insertion bias

Table 1.1: A list of insertion sequencing methods. Article, organism studied, description of sequencing library method, insertional mutagen used, and description of the screen is presented. *Note, Rinaldi et al. (2012) performed pooled sequencing of insertion sites but not as a transposon sequencing experiment.

1.3.1 Design of experiments

1.4 Association Studies in C. elegans

Only recently have the resources necessary to study the effect of natural genetic variation on on quantitative traits have been developed in *C. elegans* (Gaertner and Phillips (2010); Andersen et al. (2012, 2015); Cook et al. (2016)). A collection of 97 wild isolates (Andersen et al. (2012)) was curated and genotyped by restriction-site associated DNA sequencing (RADseq). Though the genetic diversity of *C. elegans* was reportedly low, the authors were able to map alleles with large phenotypic effect (abamectin-resistance and aversion to the pathogen *Pseudomonas aeruginosa*) through association studies.

1.5 Chapter Summary

In this thesis, I designed and we implemented transposon sequencing in *C. elegans* (Ce-TnSeq), the first such approach in an animal model. We also performed pooled GWAS for starvation survival in *C. elegans.* In Chapter 2, I used simulation-based approaches to determine the optimal design (with regards to statistical power) of Ce-TnSeq experiments. In Chapter 3, I describe our Ce-TnSeq protocol and show results from a control experiment. In Chapter 4, I used population genotyping using restriction site-associated DNA sequencing (RADseq) to perform association studies across 95 C. elegans wild isolates for starvation survival.

2

C. elegans Transposon Sequencing: Model and Design

In this chapter, I have used *Monte Carlo* simulation to describe the statistical power of, and optimally design, a Ce-TnSeq experiment. I assume the Mos1 transposon (Boulin and Bessereau (2007)) is used for mutagenesis. In previous bacterial and other single-celled TnSeq experiments, the emphasis of experimental design was to estimate the amount of replication and culture size required to saturate the target organism's genome (Bronner et al. (2016); Barquist et al. (2016); van Opijnen and Camilli (2013)). In many cases, the transformation and transposition efficiency was so high (and the culture size limits so low) that the design was trivial (van Opijnen et al. (2009)). The design of a C. elegans TnSeq experiment, however, is not trivial due to the following aspects. First, the rate of Mos1 transposition in C. elegans is quite low, less than 1 per animal (Boulin and Bessereau (2007)). Second, the induced Mos1 insertions are heterozygous in the F1 generation, requiring several generations of self-crossing to produce an adequate number of homozygous insertion mutants. Finally, our possible culture sizes are orders of magnitude smaller than cell culture, with cultures greater than 10-20 million animals becoming too cumbersome (population density requirements force these cultures to be comprised of several > 1 liter, temperature-controlled, agitated flasks). This complexity necessitates the use of simulation to explore the effect of experimental parameters on the power of a Ce-TnSeq experiment.

2.1 The Size of a Saturating Screen

The first part of designing any screen is determining the number of animals to needed to saturate the screen. The Mos1 transposon has been previously used for insertional mutagenesis in the worm, and inserts into a random TA-dinucleotide (Bessereau et al. (2001); Vallin et al. (2012)). The transposon is mobilized into the genome of F1 animals, causing them to be heterozygous for insertions. I used the empirical mutation rate of the Mos1 transposon (Vallin et al. (2012)) and *C. elegans* genome to determine the distribution of mutations in an F1 population.

Let $X = \mathbf{x}, \mathbf{x} = \langle x_1, x_2, ..., x_m \rangle$ be a random variable (r.v.) denoting the genotype of an animal where m is the number of TA-dinucleotides in the genome and $x_i \in \{0, 1, 2\}$ denotes whether the animal is homozygous wild-type, heterozygous mutant, or homozygous mutant, respectively for the ith locus.

Let $X^{(j)} = \langle \mathbf{x_1}, \mathbf{x_2}, ..., \mathbf{x_{n_j}} \rangle^T$ be the $n_j \times m$ r.v. denoting the genotypes of the jth generation, where n_j is the total number of animals in that generation. The F1 generation has several unique constraints corresponding to the behavior of the Mos1 mutagen. First, each mutation is heterozygous; formally,

$$X_{k,i}^{(1)} \leqslant 1, \forall k, i \tag{2.1}$$

Second, the number of mutations per F1 animal, $M_k = \sum_i \mathbb{1}_{X_{k,i}^{(1)} \geqslant 1}$ follows the empirically-derived Mos1 rate distribution (Vallin et al. (2012); Boulin and Bessereau (2007))(Table 2.1). Note, the expected number of mutations per animal is $\mathrm{E}[M_k] =$

Table 2.1: Empirical distribution of Mos1 mutation rate. Rates were calculated by the published rate of any mutation (Boulin and Bessereau (2007)) and a collection of over 13,000 mutants (Vallin et al. (2012)).

	$m_k = 0$	1	2	3	4	5
$Pr(M_k = m_k)$	0.5	0.484	1.56E-02	1.50E-04	3.75E-05	3.75E-05

0.52, and the expected number of mutations per *mutant* animal is $\mathrm{E}[M_k | M_k \geqslant 1] = 1.03$.

We consider a locus i affecting a particular gene g if i is within the *functional region* of g. We define the functional region of g as any nucleotide within its 1kB upstream promoter region, its 5' untranslated region (UTR), its exons, or its 3' UTR (introns are not considered functional). Let $G_i = g_i$ be the r.v. denoting the gene associated with locus i. The value of g_i is between 0 and the number of protein-coding genes in the genome (20252, genome version WS256). When $G_i = 0$, we consider the locus *non-informative* (i.e., intergenic or intronic). The Mos1 transposon inserts into a random TA-dinucleotide (Bryan et al. (1990)), therefore,

$$P(G_i = g_i) = \begin{cases} \frac{\text{\#TAs in } g_i\text{'s functional region}}{\text{\#TAs in entire genome}} & \text{for } g_i > 0 \\ 1 - P(G_i > 0) & \text{otherwise,} \end{cases} \tag{2.2}$$

I sought to determine the relationship between the number of F1 mutant animals, and the percentage of protein-coding genes with at least one mutation in the F1 generation. For a given F1 size, n_{F_1}, I simulated n_{F_1} draws from the Mos1 mutation rate distribution. For each draw, each mutation was assigned to a random TA-dinucleotide in the genome. If the TA resided within the functional region of a gene, that gene was considered hit. This allowed us to consider the distribution of gene sizes in our calculation. Figure 2.1A shows the results. We require over 400K F1 animals to hit every gene in the genome. 75K animals are required to hit 50% of genes. Note, since these draws are independent, the results are additive over replication. That

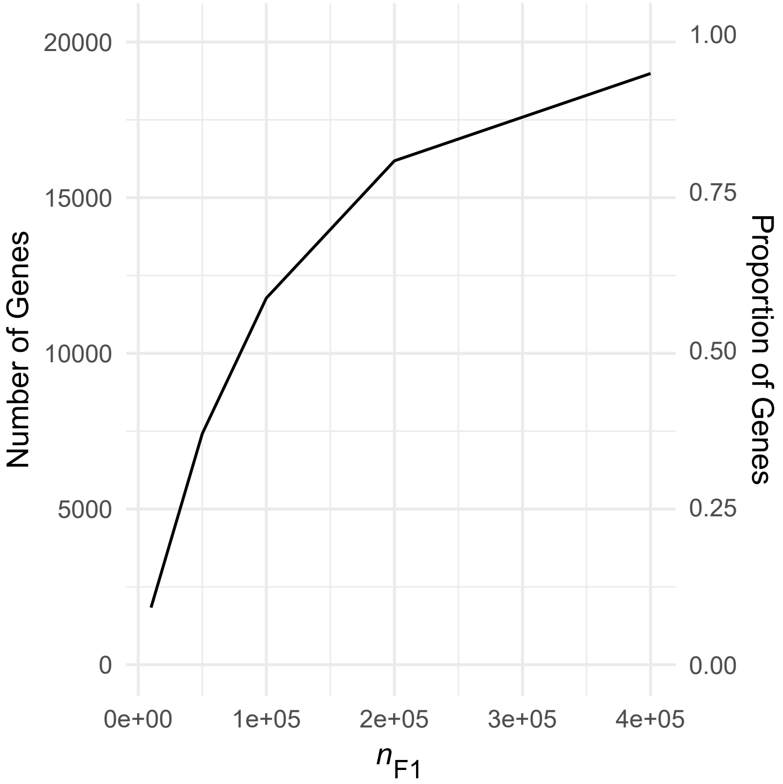


FIGURE 2.1: Number of animals to saturate genome. Using the empirical Mos1 mutation rate and the annotated C. elegans genome (WS256), I simulated populations of F1 animals. We assumed Mos1 insertions were uniform across all TA-dinucleotides. Shown is the amount of saturation expected given a particular number of F1 animals.

is, 400K F1 animals in one experiment will provide the same saturation as 100K F1 animals over four experiments. We also see a linear relationship between the number of animals and the expected number of *independent* insertions per gene (Fig. 2.1B).

2.2 Maximizing Homozygosity

We expect that transposon insertions will typically result in loss-of-function for the affected gene. Therefore, to achieve functional mutants, the heterozygous F1 generation must be self-crossed for several generations. A simple Mendelian model shows that the number of heterozygous animals decreases each generation and the number of homozygous animals increases. However there is a cost with each generation.

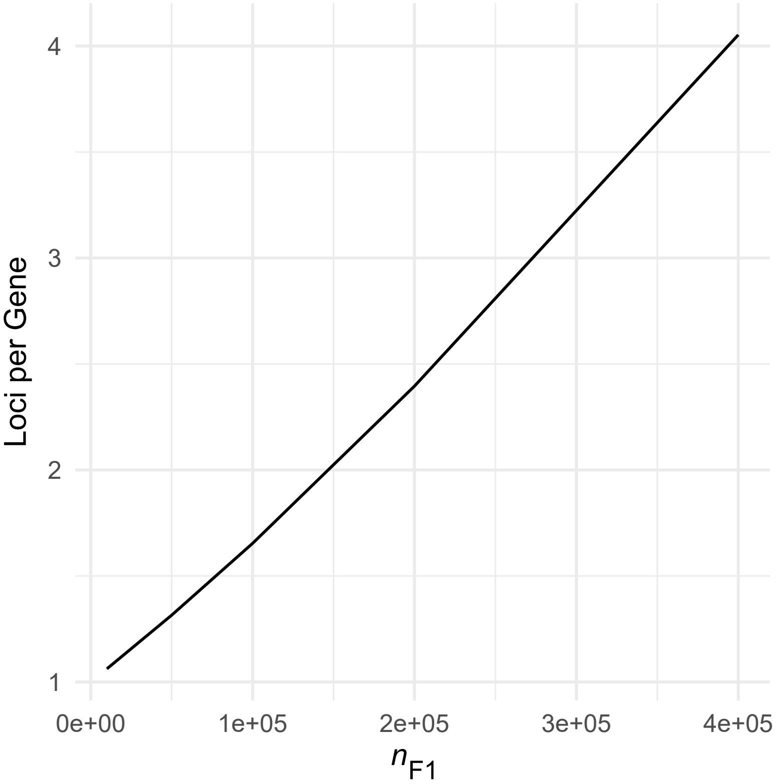


FIGURE 2.2: Independent loci per gene by number of animals. Using the empirical Mos1 mutation rate and the annotated C. elegans genome (WS256), I simulated populations of F1 animals. We assumed Mos1 insertions were uniform across all TA-dinucleotides. Shown is the expected number of different loci mutated per gene given a particular number of F1 animals.

There are the immediate concerns of labor and time (5 days per generation), but also that the exponential increase in population size between generations quickly becomes unmanageable. The result being that populations have to be sub-sampled to below the culture limit in later generations. That sub-sampling may cause mutant lineages to be excluded from the final generation.

I sought to determine the effect of F1 size, brood collection strategy, population limit, and generations of self-crossing on the expected number of homozygous animals per transposon mutant. Let Y be the m-element genotype of a child of X. We assume

Mendelian inheritance of self-crossing progeny, or formally,

$$P(Y_i = 0|X_i = x_i) = \begin{cases} 1, & x_i = 0 \\ 0.25, & x_i = 1 \\ 0, & x_i = 2 \end{cases} \tag{2.3}$$

$$P(Y_i = 1|X_i = x_i) = \begin{cases} 0, & x_i = 0 \\ 0.5, & x_i = 1 \\ 0, & x_i = 2 \end{cases} \tag{2.4}$$

$$P(Y_i = 2|X_i = x_i) = \begin{cases} 0, & x_i = 0 \\ 0.25, & x_i = 1 \\ 1, & x_i = 2 \end{cases} \tag{2.5}$$

There are several lab methods for collecting progeny from adult nematodes in liquid culture: bleach treatment of gravid worms, gravity (size) separation of adults and eggs/larvae, and optimized density preparations (i.e. setting up a culture to deplete food and arrest development at a given time/density). In my simulations, I model the number of progeny B collected from a given animal as a r.v. from a Poisson distribution with mean λ_{B_j}, where λ_{B_j} is an integer between 5 and 50, denoting the range of efficiencies for the above collection methods, to be used to collect progeny from generation j. Note, our simple Poisson model could be replaced with an empirical distribution for any of the above collection methods.

The probability of a given progeny being in the next generation depends on whether the population exceeds the max culture size. Therefore, the probability of a progeny Y being collected is:

$$P(Y \text{ being collected}) = \min(1, \lambda_{B_{j-1}} n_{j-1}/n_{\max}) \tag{2.6}$$

where n_{j-1} is the number of animals in the previous generation and $n_{\max}$ is the culture size limit.

To avoid the possible memory limitations of simulating large populations over

```
procedure SimulateGenerations($n_1, t, n_{\max}, \lambda_{B_1}...\lambda_{B_t}$)
    for $k$ in $1 : n_1$ do
        $x \leftarrow$ sampled F1 animal
        tally SimulateLineage($x, n_1, \lambda_{B_1}...\lambda_{B_t}, 2, t, n_{\max}$)
    end for
end procedure
procedure SimulateLineage($x, n_{i-1}, \lambda_{B_1}...\lambda_{B_t}, i, t, n_{\max}$)
    if $i = t$ then
        return $x$
    else
        $n_i \leftarrow n_{i-1}\lambda_{B_{i-1}}$
        $m \leftarrow \text{Poisson}(\lambda_{B_i})$
        $m' \leftarrow \text{Binom}(m, \min(1, n_{\max}/n_i))$
        $y \leftarrow \{\}$
        for $k$ in $1 : m'$ do
            $z \leftarrow$ sampled self-crossed progeny of $x$
            $y \leftarrow y \cup$ SimulateLineage($z, n_i, \lambda_{B_1}...\lambda_{B_t}, i+1, t, n_{\max}$)
        end for
        return $y$
    end if
end procedure
```

Figure 2.3: Pseudo-code for self-crossing simulation. *SimulateGenerations* is the master procedure running the simulation, and *SimulateLineage* is the recursive lineage simulator. Parameters to the simulation are: n_1 the number of F1 animals, t the number of generations to self, $n_{\max}$ the culture size limit, and $\lambda_{B_1}...\lambda_{B_t}$ the collection method mean for each generation.

time, I designed my simulation to model a single F1 lineage recursively. Pseudo-code for the simulation is shown in Figure 2.3.

I exhaustively evaluated combinations of parameters, and recorded the expected number of homozygous and heterozygous animals from a given F1 lineage in the final generation. I allowed the number of generations to vary from 3 to 5; each λ_{B_i} to independently be 5, 20, or 50; the culture limit to be either 10 million or 20 million animals; and the F1 size to vary from 10K to 50K in steps of 10K. Therefore, it was a total of 1170 parameter sets.

Of the parameters I varied, the different brood collection methods do not have an associated cost, but do affect the final counts of homozygous mutants. Therefore, I reported only the most optimal (in terms of number of homozygous animals) brood collection strategies. Table 2.2 and Figure 2.4 show the results. As expected, we see

Table 2.2: Optimal brood collection. For a given number of generations $j \in \{3, 4, 5\}$, F1 size n_{F1}, and population limit $n_{\max}$, the optimal brood collection means $\lambda_{B_1} ... \lambda_{B_4}, \lambda \in \{5, 20, 50\}$ are shown with respect to expected number of homozygous mutants per locus in F_j.

j	n_{F1}	$n_{\max}$	λ_{B_1}	λ_{B_2}	λ_{B_3}	λ_{B_4}	Mean Het.	Mean Hom.	SD Het.	SD Hom.
3	10000	20 mil.	50	50			488	765	99	164
3	20000	20 mil.	20	50			242	391	85	133
3	30000	20 mil.	20	50			167	263	53	78
3	40000	20 mil.	50	20			122	196	28	42
3	50000	20 mil.	20	50			96	154	34	50
4	10000	20 mil.	5	50	50		250	948	157	556
4	20000	20 mil.	5	20	50		132	476	82	273
4	30000	20 mil.	5	20	50		84	317	50	191
4	40000	20 mil.	5	5	50		64	248	52	148
4	50000	20 mil.	5	50	5		51	193	36	104
5	10000	20 mil.	5	5	5	20	124	1017	99	660
5	20000	20 mil.	5	5	5	50	68	513	51	301
5	30000	20 mil.	5	20	5	50	43	340	29	193
5	40000	20 mil.	5	5	5	5	34	260	27	157
5	50000	20 mil.	5	5	5	5	23	189	20	126

the ratio of homozygous to heterozygous animals to be invariant with respect to initial F1 size; as there is no selection it depends solely on the number of generations. We also find that the number of homozygous animals decreases as the F1 size increases. This is because that for most parameter sets, the population will reach the culture size limit before the last generation. Once the sub-sampling occurs due to the limit, each lineage is sub-sampled and its frequency in the population is a function of population complexity. Increasing the culture size limit increases the number of homozygous animals across parameter sets. With the size limit at 20 million, even with 50K F1 animals we can have more than a hundred homozygous mutants per lineage in the final culture.

We saw an interesting trend in the optimal brood collection strategy (Table 2.2). The negative effects of sub-sampling the population seem to be mitigated by slowly growing the population (e.g., collecting 5 eggs per adult) and collecting more eggs at the final generation; essentially delaying the generation by which the culture limit

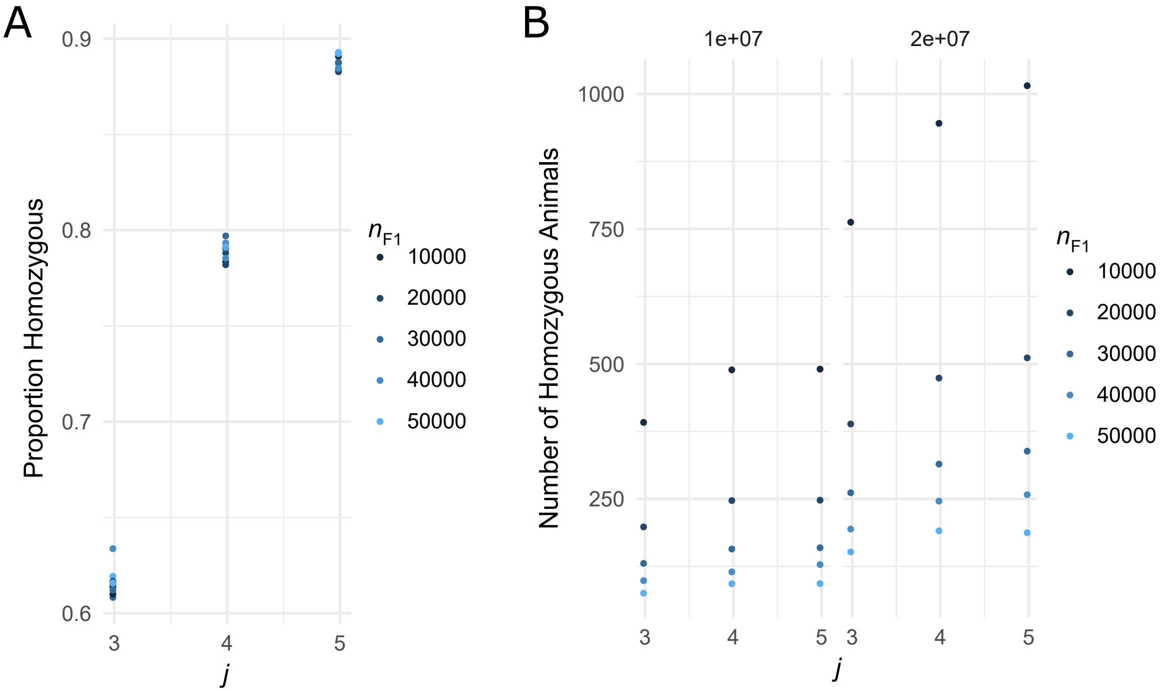


Figure 2.4: Expected number of homozygous mutants in different scenarios. A. The percentage of mutant animals (heterozygous plus homozygous) that are homozygous as a function of terminal generation. As expected, the ratio of homozygous animals increases each generation and is independent of initial culture size. B. The absolute number of of homozygous animals expected as a function of terminal generation (x-axis), initial culture size (color), and population limit (left and right panels, 10 and 20 million respectively).

is reached. Note, this result is also dependent on the distribution used to model the brood collection. The Poisson has a lower variance (equal to the mean) at low values. If the lower collection strategy has more variance (i.e. more animals have 0 progeny), one would expect fewer lineages surviving to the final generation.

2.3 Maximizing Power of a Starvation Survival Assay

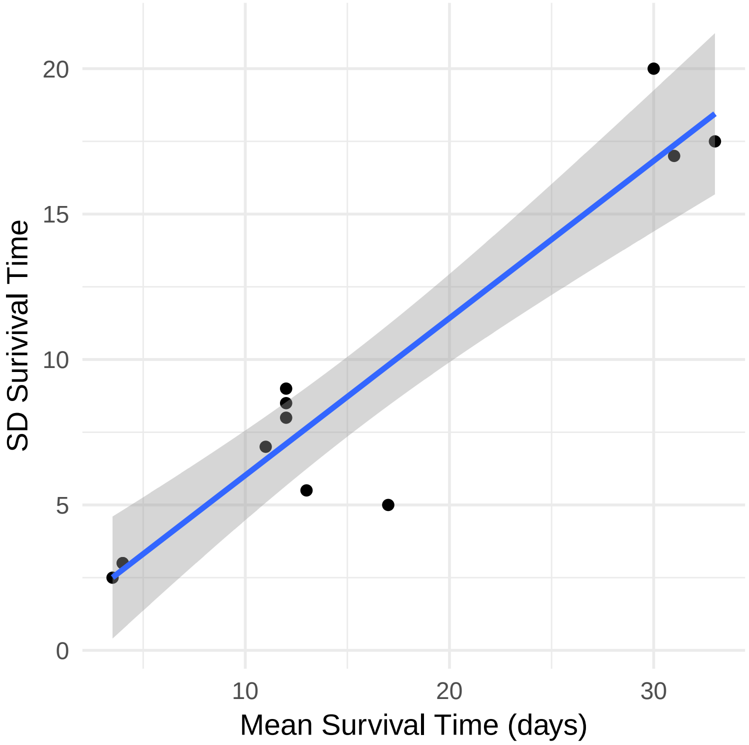
I sought to characterize Ce-TnSeq in terms of statistical power. Consider an experiment where a sample of larvae are subjected to some selection (e.g. a toxicological study, starvation, oxidative stress). The goal of such an experiment is to identify mutant loci, and their corresponding genes, that affect the ability of the animal to survive the stress (becoming either more sensitive or more resistant). The experimental population is sampled before and after treatment, and the corresponding shifts in allele frequency are measured. The statistical problem is determining which allele frequency changes are statistically significantly different from wild-type.

Our ability to detect changes in allele frequency depends on the biological variability of the stress response, the size of each sample before/after treatment, the depth of sequencing coverage, and the fidelity of the Ce-TnSeq library preparation. I extended our previous simulation to account for these different sources of variation, and investigated their consequences on our power to detect functional alleles in a hypothetical starvation survival assay.

Let $X^{(j)}$ be the population to subject to starvation, with size n_j. A sample $\dot{X}$ is taken before selection and a sample $\ddot{X}$ is take after, with samples sizes $n_{\dot{X}}$ and $n_{\ddot{X}}$ respectively. The probability that an animal is in $\dot{X}$ is $n_{\dot{X}}/n_j$ (the relative size of the sample). In order for an animal to be part of the post-selection sample $\ddot{X}$, it must not be taken in the first sample, it must survive the selection, and it must be in the portion of the post-selection population that makes it into $\ddot{X}$.

I used empirical data (Jon Hibshman, unpublished) to model the probability of an animal surviving starvation for a given amount of time t. I looked at the mean and standard deviation of starvation survival times (in buffer S-virgin) for wild-type (N2), daf-16 (a known strongly sensitive mutant), and daf-2 (a known strongly resistant mutant). The survival function for each strain was fit by a normal distribution. The means were 4, 12.8, and 31 days for daf-16, N2, and daf-2, respectively. I interpolated intermediate effects by fitting a linear regression to the standard deviation of the survival function. I assumed that mutants are recessive, so for a given animal, its survival function is wild-type if it is heterozygous, and is functionally different if it homozygous. Formally, let $S = s$ be r.v. denoting the time-of-death for given animal

FIGURE 2.5: Empirical starvation survival means and standard deviations for known mutants. Sample mean and sample standard deviations are plotted for daf-16 (sensitive), N2 (wild-type), and daf-2 (resistant) mutants (Jon Hibshman, unpublished). Blue line denotes the observed linear relationship between mean and standard deviation. Grey area denotes 95% confidence interval.



$x \in X^{(j)}$ under starvation, $P(x \text{ survives to time } t) = (1 - n_{\dot{X}}/n_j)P(S > t)$ and,

$$S \sim \begin{cases} N(\mu_{\text{mut}}, \sigma(\mu_{\text{mut}})) & , x \text{ is homozygous for mutation} \\ N(\mu_{\text{wt}}, \sigma(\mu_{\text{wt}})) & , \text{otherwise} \end{cases} \tag{2.7}$$

where t is the duration of starvation, $\mu_{\text{wt}} = 12.8$, μ_{mut} is the mutant homozygous survival time mean, and $\sigma(\mu) = 0.61 + 0.54\mu$ (the fit linear model).

The probability of a surviving animal being captured in the second sample $\ddot{X}$ is the ratio between the sample size and the number animals alive at the collection time t. We assume that the majority of mutations will have no effect on survival, therefore the size of the population at time t is $(n_j - n_{\dot{X}})P(S > t)$.

Unfortunately, we can't measure the number of animals in each sample directly.

Instead, animals are lysed, their genomes are mixed, a sequencing library is created, and that library is sequenced on a sequencer. For our analysis we will not be modeling the library process (such an effort is better suited to empirical measurement), and instead will assume the library protocol is ideal. That is, it creates an infinite pool of genome molecules that maintains the original allele frequency of the sample. Such an assumption allows us to define a base measure of power for any Ce-TnSeq library.

Assuming each sample is prepared ideally, we can model the sequencing noise with a Poisson distribution. Note, in RNAseq models a negative binomial is preferred to the Poisson as real RNAseq data is over-dispersed (Yu et al. (2013)). I maintain that the Poisson is appropriate here as we are modeling an ideal library as well as capturing much more variability in other parts of our model. For a given locus i, let $A_i = \sum_{x \in \dot{X}} x_i$ be the total number of mutant allele counts at locus i in sample $\dot{X}$ (counting singly the heterozygous animals and doubly the homozygous). The expected number of sequenced reads at locus i, R_i, is modeled by,

$$R_i = r_i | A_i, A \sim \text{Poisson}(\frac{A_i}{A} n_r) \tag{2.8}$$

, where $A = \sum_{x \in \dot{X}, 1 \leqslant i \leqslant m} x_i$ is the total number of tranposon alleles in the sample and n_r is the total number of reads.

With the above model, I sought to explore the statistical power we have to detect loci with varying magnitudes of effect on starvation survival. The power, or $1 - $ probability of Type II error , is the probability of detecting a significantly different locus i, given a p-value threshold of α. Using my previously calculated optimal parameters for creating homozygous $X^{(j)}$ populations, I varied the time of starvation t (8, 12, and 21 days), the sample size for $\dot{X}$ and $\ddot{X}$ (1-4 million live animals), and the number of reads n_r (multiplexing a 250 million read Hiseq 2500 lane by 1, 12, 24, 48, and 96). For each parameter set, I estimated the population size at t and simulated the null model (loci with $\mu_{\text{mut}} = \mu_{\text{wt}}$). Once I established the null model, I simulated mutant loci with varying magnitudes of effect (μ_{mut}), and calculated the overlap of their read distributions with the null model a p-value cutoff of 0.001. The hypotheses used in my power analysis were:

$$H_0 : \log(R_{0,\ddot{X}}/R_{0,\dot{X}}) = \log(R_{i,\ddot{X}}/R_{i,\dot{X}}) \tag{2.9}$$

$$H_{\text{a}} : \log(R_{0,\ddot{X}}/R_{0,\dot{X}}) \neq \log(R_{i,\ddot{X}}/R_{i,\dot{X}}) \tag{2.10}$$

where $R_{0,\ddot{X}}$ and $R_{i,\ddot{X}}$ are the reads from a phenotypically wild-type mutant and the mutant under consideration, respectively, after starvation.

Figure 2.6 shows the optimal power we can achieve as a function of F1 size and mutant effect (μ_{mut}). As the initial F1 culture size increase, we know from our previous results that the number of homozygous mutant animals per locus decreases. We saw that this caused a decrease in power. I also show, as expected, that our power to detect the a mutant's functional effect decreases as that effect becomes closer to the wild-type average (12.8 days).

I further showed (Figure 2.7) that our power to detect mutant effect is dependent on the starvation time relative to the mutant survival time. That is, at 8 days of starvation, we have virtually no power to detect long-lived mutants as the wild-type animals are still mostly alive at 8 days. However, we still have power to detect short-lived mutants, as their numbers will be dramatically reduced with respect to wild-type. If we increase the starvation time to 21 days, we gain the ability to detect mutants with mean survival greater than 16 days, though mutants close to the average survival (12.8 days) are still undetectable within our parameters. Note, we still maintain the ability to detect short-lived mutants after 21 days of starvation. This is due to their still being a detectable number of wild-type animals alive at day 21. We expect our ability to detect short-lived mutants to dramatically decrease at longer starvation times.

I also assessed the impact of sample collection size (e.g. $n_{\dot{X}}$ and $n_{\ddot{X}}$, for simplicity,

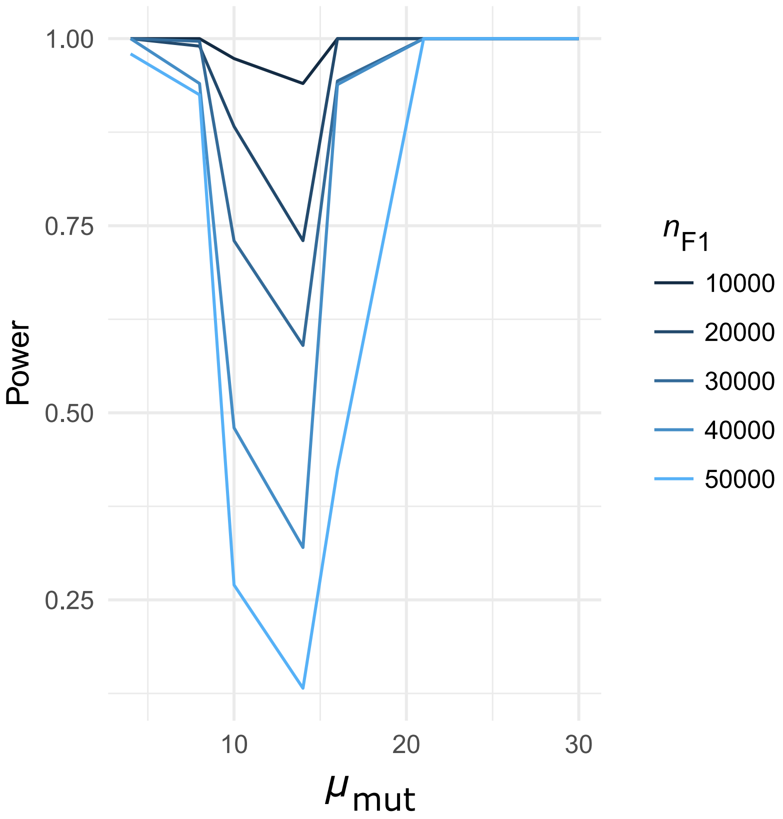


FIGURE 2.6: Simulating the power of a starvation survival Ce-TnSeq experiment. Optimal power is plotted as a function of mutant mean survival (μ_{mut}, x-axis) and initial culture size (color).

$n_{\dot{X}} = n_{\ddot{X}}$) on power in Figure 2.8). For small F1 culture sizes where we have an ample amount of homozygous mutants, we see diminishing returns on increasing the sample collection size past 2 million animals. For the largest F1 culture size (50K) with the fewest number of homozygous animals, we actually see a drop in power for larger sample sizes. This is due the $\dot{X}$ sample taking too many mutant animals and reducing the number of animals post-starvation. We also saw an unexpected increase in power from 30K to 40K F1 size. I therefore investigated the empirical distribution of the test statistic (the difference between mutant log ratios, Figure 2.8B). The uneven nature of the distributions shows the noise in calculating the distribution given the number of iterations in our simulation (300 per parameter set). We believe that this

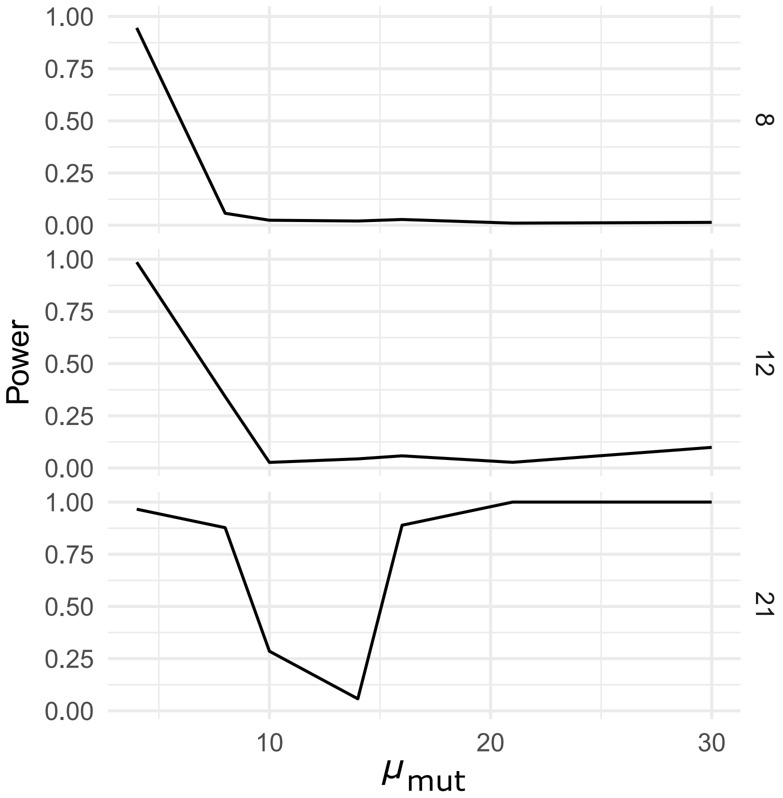


FIGURE 2.7: The effect of starvation time on power. We show how are ability to detect mutant effect varies as a function of starvation time (each panel, 8, 12, and 21 days) relative to mutant mean survival (x-axis). The data is shown for an F1 size of 30K, a terminal generation of 4, a p-value cutoff of 0.001, and optimal parameters for sample size and sequencing depth.

noise explains the unexpected order and crossover in power shown in Figure 2.8A.

Finally, I investigated the effect of sequencing depth on power (Figure 2.9). We saw that for the parameters tested (F1 size of 30K, terminal generation of 5, mutant survival of 16 days, starvation time of 21 days, p-value cutoff of 0.001), that power was only affect after multiplexing greater than 48, or less than 5 million reads.

2.4 Conclusions

I used simulation-based analyses to optimally design a Ce-TnSeq experiment and to characterize our statistical power in a proposed starvation survival experiment. I

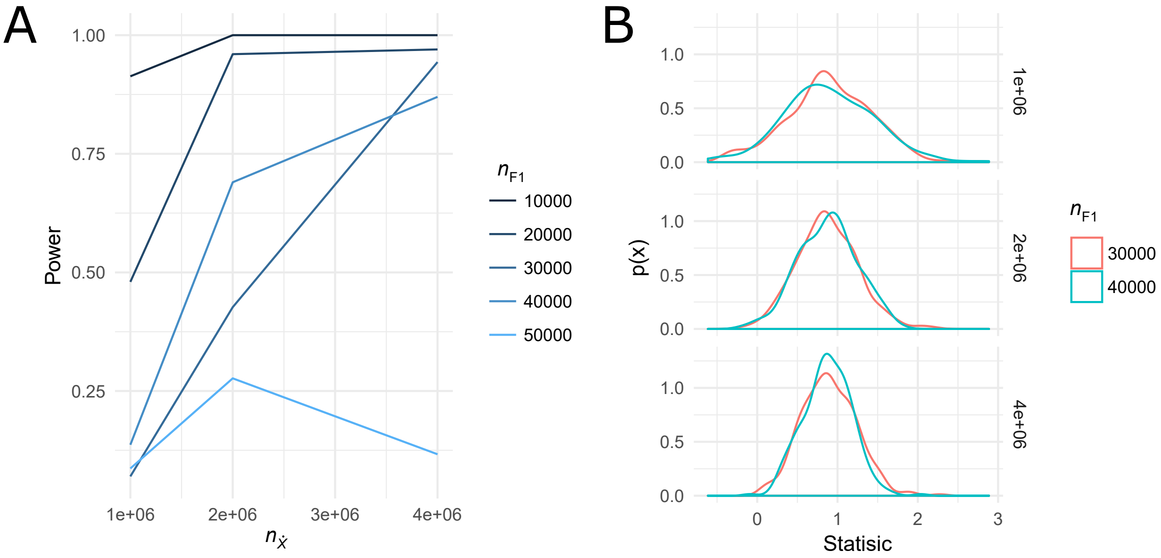


FIGURE 2.8: The effect of sample size on power. A. We show how power varies as a function of sample collection size ($n_{\dot{X}} = n_{\ddot{X}}$, x-axis) and F1 culture size (color). The data is shown mutant survival time of 16 days, terminal generation 3, population cap of 20 million, starvation time of 21 days, and 5.2 million reads. B. To investigate the cross-over between the 30K and 40K F1 sizes in panel A, we show the empirical distribution of the test statistic.

characterized the effect of various experimental parameters on the number of genes screened and the number of homozygous and heterozygous mutant animals per insertion locus. We found that a brood collection strategy that collects few animals in the first generations and collects as many progeny as possible in the final generation is optimal in terms of number of homozygous mutants per locus.

My power analysis considered many parameters: the above experiment design parameters (F1 size, brood collection strategy, population limit, number of generations), length of starvation, sample collection size, sequencing depth, and p-value cutoff. We found that at short starvation times, we have the power to detect starvation sensitive mutants but lack the power to detect resistant mutants. I found that at late time points, we had the power to detect both sensitive and resistant mutants, but postulated that too long of starvation will decrease our ability to detect the most sensitive mutants. I showed that sample collection size must be relative to the

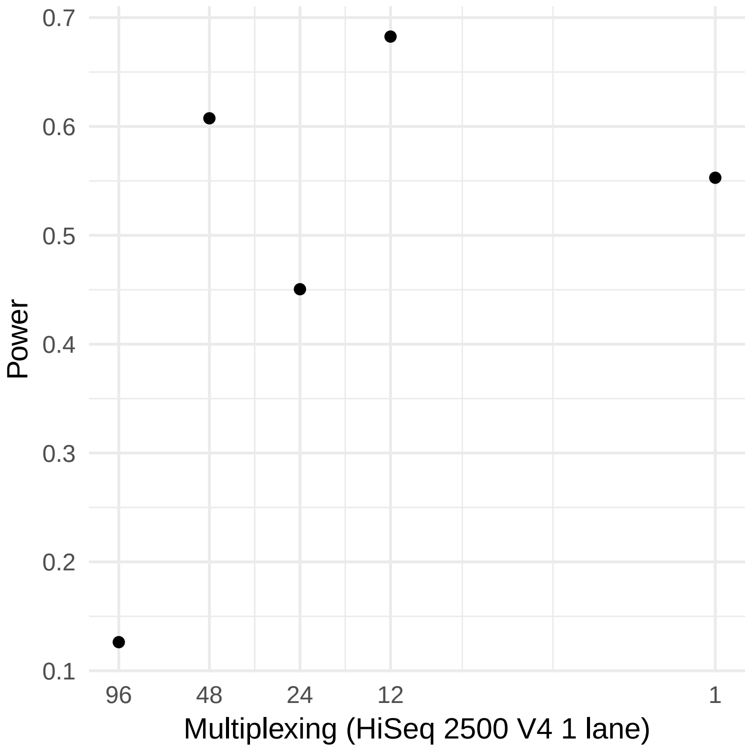


FIGURE 2.9: The effect of sequencing depth on power. Sequencing depth is represented as the degree of multiplexing on a HiSeq 2500 V4 lane (250 million reads). The data shown is for a F1 size of 30K, terminal generation of 5, mutant survival mean of 16 days, starvation time of 21 days, p-value cutoff of 0.001.

number of homozygous animals in the final generation; too few animals collected reduces power, but too many animal collected leads to diminishing returns and a drop in power when the number of homozygotes is small. Finally, I measured the effect of sequencing depth on our statistical power, and showed a that a high level of multiplexing (48 libraries) is possible with the parameters tested.

To our knowledge, this is the first design analysis of a transposon sequencing experiment that considers diploidy. We plan to use our simulation to design our actual Ce-TnSeq experiments. The simulation is also invaluable in identifying potential areas for improvement in the molecular biology of Ce-TnSeq. For example, large gains in saturation and power can be had if the mutation rate can be increased, or homozygous mutants be selected for in the final generation.

As is, our simulation approximates an ideal Ce-TnSeq protocol by modeling read counts with a Poisson distribution. It would be straightforward to replace this distribution with empirical data from a Ce-TnSeq library. This would allow us to anticipate the power of future Ce-TnSeq experiments.

3

C. elegans Transposon Sequencing (Ce-TnSeq)

Transposon sequencing is a method for measuring the effect of gene loss-of-function at a genome-wide scale (van Opijnen et al. (2009)). It combines transposon-mediated mutagenesis and insertion site sequencing. A saturated transposon mutant population is created and subjected to selection (or stratification) for a particular phenotype. Mutant allele frequencies in the population are measured before and after selection, and the functional effect of the gene loss-of-function is inferred from the allele frequency shift.

To date, transposon sequencing has been limited to single-celled microorganisms or cell culture (Table 1.1). Here, we implemented transposon sequencing in the metazoan model, *Caenorhabditis elegans*. Having transposon sequencing in the nematode would allow us to perform genome-wide characterizations of gene function for a variety of phenotypes. We will be able to screen tissue-specific and hormone-directed pathways, such as insulin-like signaling.

Transposon-mediated mutagenesis has been previously done in *C. elegans* (Bessereau et al. (2001)) using the exogenous *mariner* transposon Mos1 (Bryan et al. (1990)).

The existing method relies on a transgenic mutator strain which contains the Mos1 transposon and the mobilizing transposase enzyme (under a heat shock-inducible promoter) on two separate extra-chromosomal arrays. Young reproductive mutator adults are heat-shocked, causing the transposase to mobilize the transposon elements in the array into the genome. This occurs in both the soma and the distal germline. The germline insertions cause the F1 progeny of the heat shocked adults to be heterozygous for transposon insertions (around 50% of F1 animals contain 1-2 insertions). The F1 animals are then either screened immediately or self-crossed for a single generation (F2).

The method of Mos1 transposition is shown in Figure 3.1. The transposase excises the Mos1 transposon (causing a double-stranded break) and inserts it into a random TA-dinucleotide. Previous work (Vallin et al. (2012)) has concluded that the insertion event is relatively unbiased across the genome (over 13,300 mutant animals considered).

There are several challenges in adapting the existing Mos1 mutagenesis method for transposon sequencing. First, the current method requires manual or large particle fluorescence-based sorting to maintain the doubly-transgenic mutator strain (the separate extra-chromosomal arrays for the transposon and transposase have a 30%-40% percent chance of degrading between generations). Second, the mutant population itself must be sorted against the transposon-containing array otherwise the genotyping would only report the array location (the transposon is at high copy number, 300-800, in the array). Finally, the existing method uses a restriction enzyme-based inverse PCR method. This method can only detect insertions that are at the right distance from the associated cut site. The inverse PCR is also designed around Sanger sequencing, and as is does not result in a next gen sequencing library.

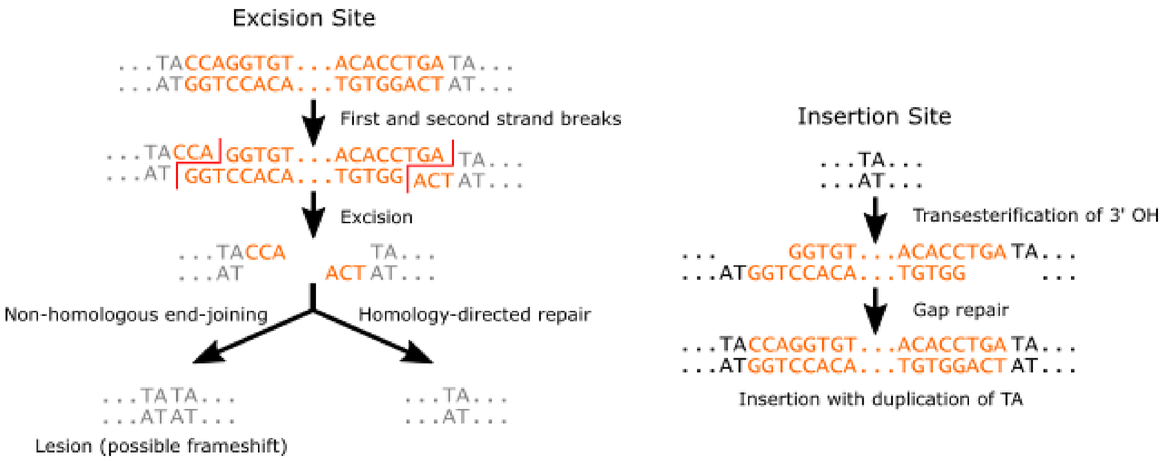

Excision Site
...TACCAGGTGT...ACACCTGATA...
...ATGGTCCACA...TGTGGACTAT...
First and second strand breaks
...TACCA GGTGT...ACACCTGA TA...
...AT GGTCCACA...TGTGG ACT AT...
Excision
...TACCA TA...
...AT ACT AT...
Non-homologous end-joining
Homology-directed repair
...TATA...
...ATAT...
Lesion (possible frameshift)
...TA...
...AT...
Insertion Site
...TA...
...AT...
Transesterification of 3' OH
... GGTGT...ACACCTGA TA...
...ATGGTCCACA...TGTGG ...
Gap repair
...TACCAGGTGT...ACACCTGATA...
...ATGGTCCACA...TGTGGACTAT...
Insertion with duplication of TA

FIGURE 3.1: The method of Mos1 excision. Orange denotes the Mos1 transposon sequence. Light grey denotes flanking sequence from the excision site. Black denotes genomic flanking sequence at the insertion site. At each inverted terminal repeat (ITR) a Mos1 transposase causes a staggered break and forms a paired end complex (PEC). The PEC uses the 3 OH- either end to insert into a TA-dinucleotide. Endogenous gap repair fills in and completes ligation at the insertion site. The remaining double-stranded break at the excision site is closed by endogenous DNA repair machinery. If non-homologous end-joining (NHEJ) is used, a lesion of in the form of the duplicated TA may remain. This can cause a frameshift if the excision site is a coding region.

3.1 Redesigning the Mos1 Mutagenesis System

We sought to increase the throughput and reduce the labor of the existing Mos1 mutagenesis system. We therefore engineered a new mutator strain, LRB311 (dukEx125[Prps-0::hygR+Pmyo-2::mCherry+Phsp16.48::MosTase+Mos1 AscI Substrate]), that addressed the drawbacks of the original. First, I combined the transposon and transposase onto a single extra-chromosomal array. This allows us to select for only a single marker for the mutator strain. Second, in addition to a single fluorescent marker (allowing for manual screening), I inserted a recently developed hygromycin resistance marker (Tas et al. (2015)). This drug-based marker is easy to use, cheap, and can be added to normal culture conditions. Finally, I engineered a rare restriction site on either flank of the transposon in the extra-chromosomal array (Figure 3.2). This allows us to avoid removal of transposon array-containing animals prior

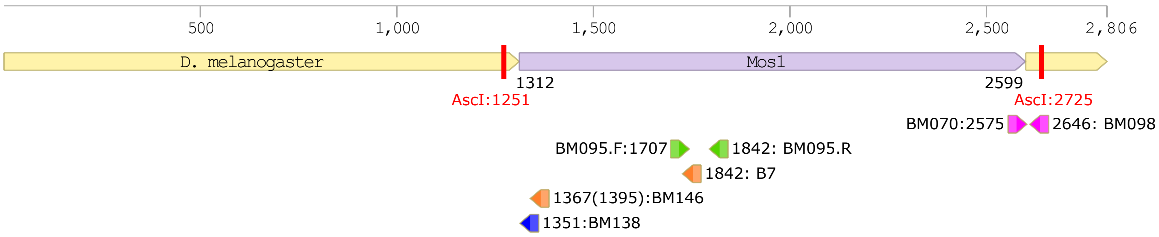


Figure 3.2: Insert map for pBM020. pBM020 contains the transposon DNA (purple) for the mutator strain. AscI cut sites (red line) allow for the transposon to be selectively degraded within the extra-chromosomal array. PCR primers BM070/BM098 (magenta) are used to detect array contamination in mutant lines. PCR primers BM095.F/R are used to detect the presence of the transposon. Primers B7 and BM146 (orange) are used in the nested PCR of the Ce-TnSeq protocol. In parenthesis is the location of BM146 including the overhang portion of the primer. BM138 (blue) is the custom read 1 sequencing primer. See supplemental data for complete plasmid map. The above insert was cloned into the multiple cloning site of PUC19. See supplemental data for full plasmid map and sequence.

to genotyping, as we can selectively digest the array transposons in our DNA prep.

We tested the mutagenic rate of strain LRB311 as follows. We heat-shocked 10 young adults (P0) according to the existing protocol (see Materials and Methods). We randomly selected and isolated 27 F1 progeny from the P0. The F1 progeny were self-crossed for 2 generations (F3). DNA was extracted from 10-12 array-negative (based on lack of fluorescence) animals per F1. This DNA split was into 3 and PCR-amplified for: genomic DNA (daf-16, positive control), the extra-chromosomal array, and transposons (array or genomic). Figure 3.3 shows the results. In our single replicate, we observe 78% of animals containing a transposon insertion. This number agrees with the roughly %50 rate reported for the existing method (Boulin and Bessereau (2007); Vallin et al. (2012)). We also observe animals that do not have a transposon, suggesting that our strain is free from genomic insertions (which would be present in all derived mutants).

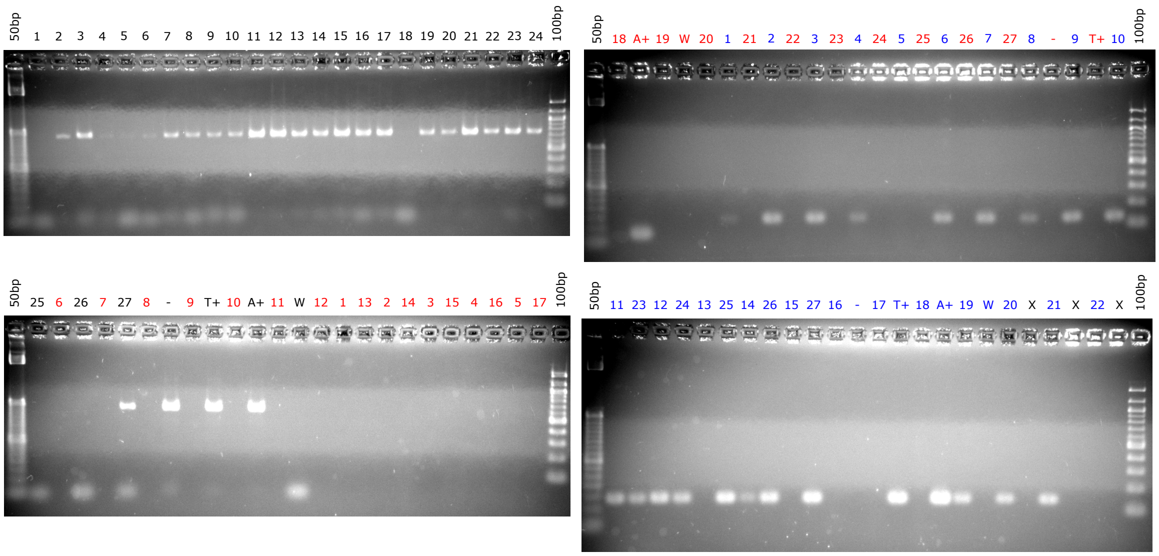

50bp
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24
100bp
50bp
18 A+ 19 W 20 1 21 2 22 3 23 4 24 5 25 6 26 7 27 8 - 9 T+ 10
100bp
50bp
25 6 26 7 27 8 - 9 T+ 10 A+ 11 W 12 1 13 2 14 3 15 4 16 5 17
100bp
50bp
11 23 12 24 13 25 14 26 15 27 16 - 17 T+ 18 A+ 19 W 20 X 21 X 22 X
100bp

FIGURE 3.3: Mutation rate of Mos1. 1.5% agarose gel (Ethidium Bromide). 1-27 are the 27 F1 animals screened, - is the wild-type N2 negative control, T+ is a single-copy transposon insertion strain positive control (IG129), A+ is a extra-chromosomal array positive control (LRB311), and W the water control. In black, gDNA positive primers (daf-16, expected length 335bp); in red, array-specific primers (BM070/BM098, expected length 73bp); and in blue, transposon-specific (genomic or array) primers (BM095, expected length 135bp). Our controls behave as expected, and with censoring no-band gDNA samples (1,18,25-26) we observe a mutation rate of 78%. (Experiment performed by Rojin Chitrakar)

3.2 Creating a *C. elegans* Transposon Sequencing Library

After developing LRB311, a Mos1 strain capable of high-throughput mutagenesis, I sought to design an insertion site sequencing protocol for the Mos1 transposon. The existing method for identifying Mos1 insertions in *C. elegans* uses an restriction site-dependent inverse PCR reaction (Boulin and Bessereau (2007)). This method only identifies insertion sites that are a certain distant from a restriction site, which leads to many possible Mos1 sites being undetectable.

A variety of approaches have been applied to other transposons and other organisms (see Table 1.1 for a summary). We were restricted in the approaches we could consider as the Mos1 transposon is sensitive to internal modification as it reduces its insertion rate (Lohe and Hartl (2002)). I decided on a PCR-based approach for selective amplification and sequencing library adaptation of the insertion sites. This approach is promising because it requires only minor modifications to a commercial Illumina kit (NEB DNA Ultra, Cat. E7370S), and a conceptually similar approach has been successfully applied to *plasmodia falciparum* (Bronner et al. (2016)).

Our insertion site sequencing method for Mos1, *C. elegans* transposon sequencing, or Ce-TnSeq, is as follows. First, genomic DNA from a mutant population is isolated and sonicated (Figure 3.4). This DNA will contain transposons inserted into the genome as well as transposons within the extra-chromosomal array (the relative frequency of each depends on the generation the DNA was harvested as the array degrades between generations). This sonicated DNA is then end-repaired and A-tailed according to the NEB DNA Ultra kit's specifications. Then, according to the kit protocol, Y-tailed adaptors are ligated either end of each A-tailed DNA fragment (including transposon-less gDNA). The design of these primers prevent a single primer from amplifying an adapted fragment. At this point, I introduce an additional step: we digest the ligated DNA fragments with AscI, a rare-cutting restriction enzyme whose sites are engineered adjacent to array transposons (Figure 3.4). This digest removes the adaptors from array transposon fragments, effectively silencing them in the further steps of the protocol.

The NEB kit uses a bead-based method for selecting ligated fragments of a specified size (as well as removing unligated adaptors from the prep). The typical recommended size for fragments is 200bp; I increased this to 500bp as we intend to sequencing from the flank of the transposon, and that flank can occur anywhere in the initial sonicated DNA fragments.

Typically, primers designed for each side of the Y-tailed adaptors are then used to amplify the library and append Illumina-specific sequence to each fragment. However, we wish to only amplify the fragments containing transposon insertions. Therefore, we use only one adaptor PCR primer, and replace the other with a primer specific to the left flank of the transposon (Figure 3.5). For additional specificity, I designed this PCR reaction to be a nested reaction, requiring two separate PCRs with separate primer pairs. This additional specificity was shown to be required by our work (results not shown) and in similar protocols (Bronner et al. (2016)). Finally, we use a custom sequencing primer (BM138) designed to anneal precisely at the the leftmost flank of the transposon. This allows us another degree of specificity, as well avoiding degenerate (i.e. from the transposon) sequence in the read 1 reads.

It should be noted that this protocol can also be extended to capture the right flank of the transposon. As the transposon itself is asymmetric (there are even minor differences in the ITRs), the adaptor-ligated sample would need to be split into two separate nested PCR reactions (one for each flank). The separate libraries could be combined prior to sequencing, but an additional custom sequencing primer would need to be designed for the right flank. The benefit of sequencing both flanks would be additional specificity (putative insertion loci that did not have reads from both flanks could be censored) as well as controlling for sequence-specific bias in the read counts (read counts for either flank of a specific locus could be averaged).

3.3 Sequencing Results for Ce-TnSeq Control Libraries

To test our Ce-TnSeq method, we made a panel of six control libraries: N2 - Mos1 transposon-less genomic DNA from the standard lab strain, N2+PCR - the previous DNA with an addition of PCR product from the transposon flank in the array, Standard - a controlled mixture of DNA from 15 different Mos1 transposon mutant strains (see Materials and Methods), IG129 - gDNA from a single Mos1 mutant strain, Array - gDNA from LRB311 (should only have the transposon in the array), and F1 - a Mos1 mutant population of heat shocked LRB311 parents' progeny (should contain genomic and array insertions). All samples were prepared using the Ce-

TnSeq protocol, except the AscI digest was only performed on the F1 sample. N2 is our negative control to rule out any non-specific amplification of gDNA. N2+PCR is our positive control which should also show a single band in the library prep (other controls will show smears due to sonication). Standard contains 3 groups of 5 Mos1 mutant strains (15 total) combined over a 3-fold scale; it is a positive control that should allow us to measure are sensitivity at identifying insertions at different genomic locations at different concentrations. Array is another positive control that should allow us to confirm the flanking sequence of the transposon in the array. Finally, F1 is our experimental condition; we should be able to recapitulate the expected Mos1 mutation rate from the population as well as measure any site-specific bias in our amplification (as the F1 population is heterozygous and the insertions are independent, each insertion should be at the same frequency).

An agarose gel of our Ce-TnSeq control libraries is shown in Figure 3.6. We saw the expected size band (407bp) of the PCR product in N2+PCR. We also saw a similar-sized band (but fainter) in our N2 negative control, evidence of cross-contamination of the PCR product, a known issue we had in the prep (data not shown). However, contaminated N2 sample (and the N2+PCR sample) showed no visible other product, providing evidence for specific amplification of insertion sites only. The remaining samples showed multiple banding patterns and smears, evidence that multiple sonicated fragments were being amplified by the protocol. Note, the gel patterns were qualitatively similar to what was seen in a Bioanalyzer analysis of the the samples (data not shown).

We proceeded with sequencing the six samples. N2, N2+PCR, Standard, IG129, Array, and F1 were combined (multiplexed using NEB index primers) at ratios of 10%, 10%, 20%, 10%, 10%, and 40% respectively (percentages were based on how degenerate the libraries were expected to be). The combined samples were sequenced using a 150bp paired-end MiSeq Nano kit (1 million paired reads expected). Due

Table 3.1: Summary of Ce-TnSeq control read counts and duplicates. Unique locations were determined by grouping read 1 reads that were within a 25bp window of each other. Total paired reads is the total number of reads that were called *proper* by the bowtie2 algorithm (facing each other and within 1Kb apart). Total unique paired reads is the total number of paired reads that have unique read 2 read 5' locations.

Library	Total	Paired	Unique Paired	Unique Loci	Proportion pBM020
Array	101214	67650	557	213	0.99
F1	171914	79886	374	213	0.53
Standard	60676	54188	822	128	0.05
IG129	69773	35645	1097	244	0.02
N2	21021	16153	105	32	0.99
N2 + PCR	37666	28631	105	19	0.99

to the low complexity of the library, and some error in determining the libraries' molarity, approximately 462K reads passed the Illumina filter. Those reads were then truncated to 25bp and aligned to the *C. elegans* genome and the transposon array (i.e., injection plasmid pBM020) using bowtie2. Of those 462K reads, 282K (or 61%) were determined to be *proper* pairs (facing each other and within 1kb) that aligned to the genome or array (Table 3.1).

I then grouped the read pairs by the location of read 1. In our protocol, the 5' end of read 1 should identify the exact location of the transposon insertion. In general, we saw relatively few unique loci in our library (Table 3.1). In the N2 and N2+PCR samples, this is reassuring. We saw very few (10-30) non-insertion sites being detected, and the overwhelming majority of reads ($> 99\%$) came from the contaminating and spiked array PCR product. For the single insertion strain, IG129, we see a majority of reads ($> 50\%$) coming from the annotated insertion site on chromosome IV (Figure 3.7). We also see over 20% of reads coming from a single location on chromosome III. Note, this same location is found in extremely low concentrations in other samples ($< 0.001\%$). This could be a background site that was overamplified by PCR or it could be an additional insertion site not identified by the inverse PCR method used to genotype the strain. We will design PCR primers around this location to confirm this. However, we conclude that non-specific genomic amplification is not a concern across the samples.

Unexpectedly, only a relatively few number of unique loci were detected in the F1 sample. Given the previously measured mutation rate of LRB311 and the number of F1 animals in the sample (100K), we expected over 50K insertion sites in the library. One explanation could be that the mutation rate of LRB311 is highly variable and sensitive to the mutagenesis conditions. To investigate this, we plan on sequencing two additional independent F1 samples we've collected. We will also replicate the small-scale mutation rate experiment in Figure 3.3.

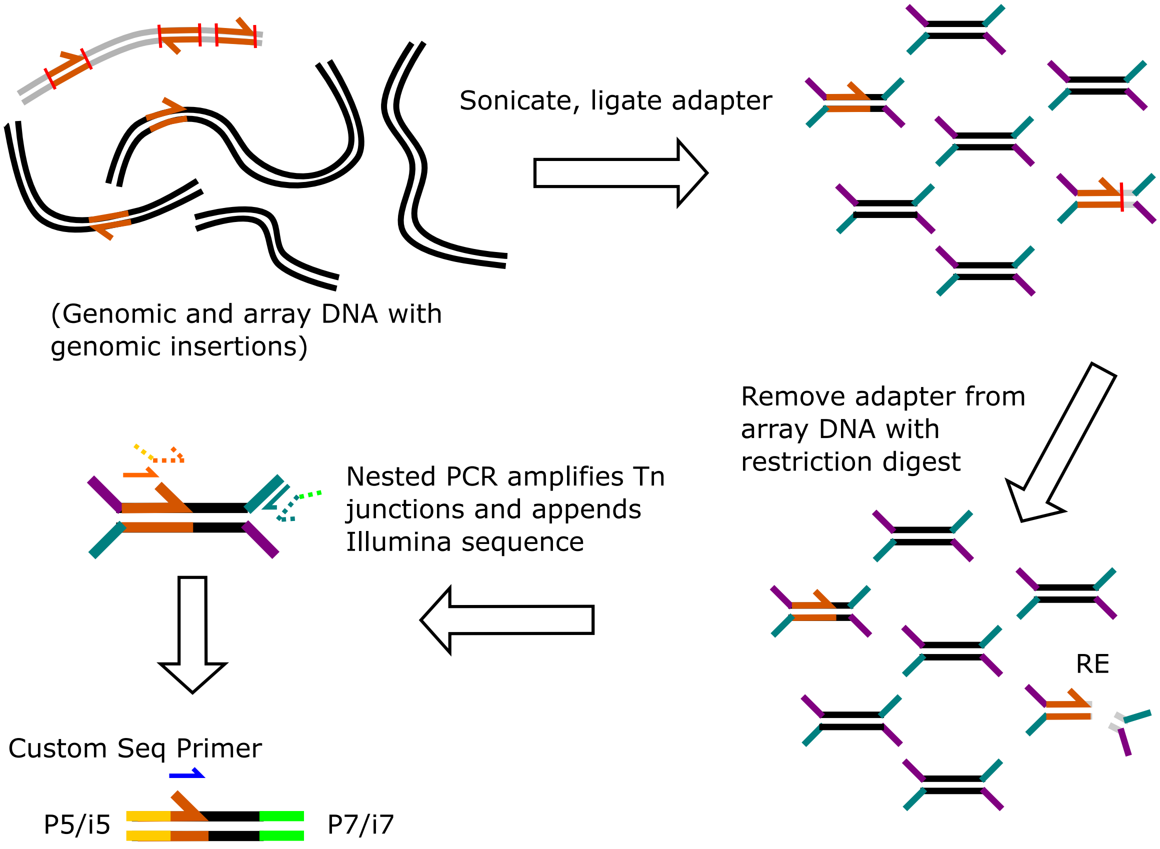

Sonicate, ligate adapter
(Genomic and array DNA with genomic insertions)
Remove adapter from array DNA with restriction digest
RE
Nested PCR amplifies Tn junctions and appends Illumina sequence
Custom Seq Primer
P5/i5
P7/i7

Figure 3.4: Overview of the Ce-TnSeq protocol. Genomic DNA (black) and array DNA (light grey) from a mutant population contains transposon insertions (orange). The transposon itself is directional (orange arrow) and inserts into any TA dinucleotide. Transposons in the array are engineered with flanking AscI cut sites (red) in order to selectively degrade them. DNA is sonicated, end-repaired, and then ligated to NEB Illumina adaptors (cyan/purple Y). The ligated DNA is subject to AscI restriction digest (RE) to cleave adaptors from extra-chromosomal insertions. Primers B7 (solid-orange) and BM144(solid-cyan) are used in PCR1 of a nested PCR reaction. PCR2 uses primers BM146 (dotted-orange) and a NEB index primer (dotted-cyan).

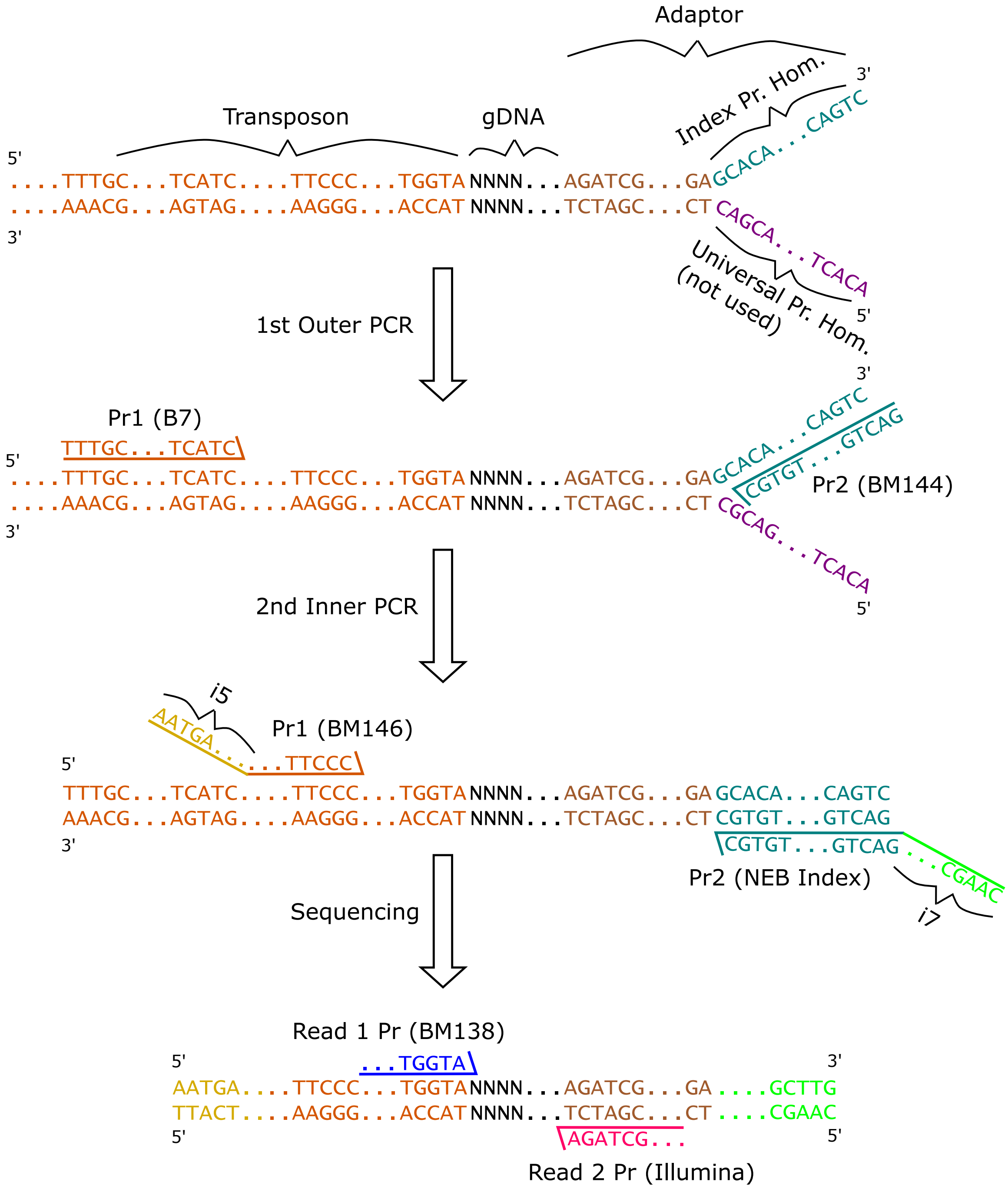

Adaptor
Transposon
gDNA
Index Pr. Hom.
3'
5'
....TTTGC...TCATC....TTCCC...TGGTA NNNN...AGATCG...GA GCACA...CAGTC
....AAACG...AGTAG....AAGGG...ACCAT NNNN...TCTAGC...CT CAGCA...TCACA
3'
Universal Pr. Hom. (not used)
5'
1st Outer PCR
3'
Pr1 (B7)
TTTGC...TCATC
5'
....TTTGC...TCATC....TTCCC...TGGTA NNNN...AGATCG...GA GCACA...CAGTC
....AAACG...AGTAG....AAGGG...ACCAT NNNN...TCTAGC...CT CGCAG...TCACA
CGTGT...GTCAG
Pr2 (BM144)
3'
5'
2nd Inner PCR
i5
AATGA.....TTCCC
Pr1 (BM146)
5'
TTTGC...TCATC....TTCCC...TGGTA NNNN...AGATCG...GA GCACA...CAGTC
AAACG...AGTAG....AAGGG...ACCAT NNNN...TCTAGC...CT CGTGT...GTCAG
3'
CGTGT...GTCAG...CGAAC
Pr2 (NEB Index)
i7
Sequencing
Read 1 Pr (BM138)
...TGGTA
5'
3'
AATGA....TTCCC...TGGTA NNNN...AGATCG...GA ....GCTTG
TTACT....AAGGG...ACCAT NNNN...TCTAGC...CT ....CGAAC
5'
5'
AGATCG...
Read 2 Pr (Illumina)

FIGURE 3.5: Details of Ce-TnSeq nested PCR. Rounds of the nested-PCR are depicted. The outer PCR (PCR1) selectively enriches the transposon-containing fragments above the genomic background. PCR2 provides additional specificity, and creates amplified product containing the required Illumina sequence for sequencing (P5 mustard, P7 bright green). A custom sequencing primer (BM138, blue) ensures that read 1 reads start immediately adjacent to the transposon.

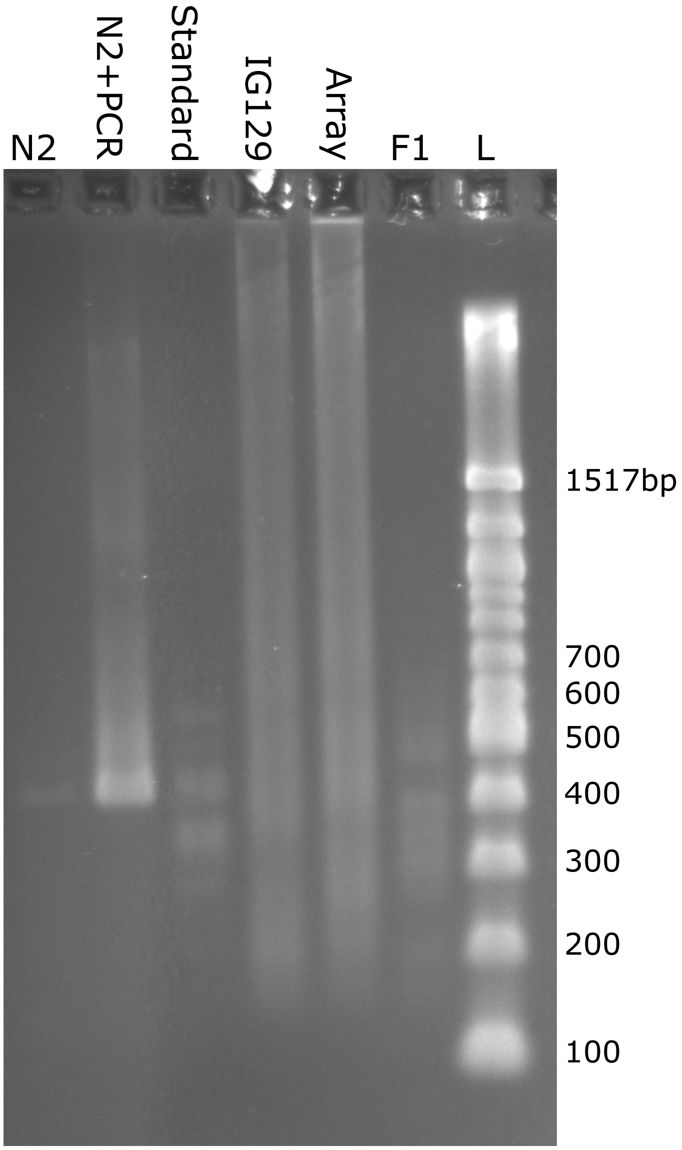
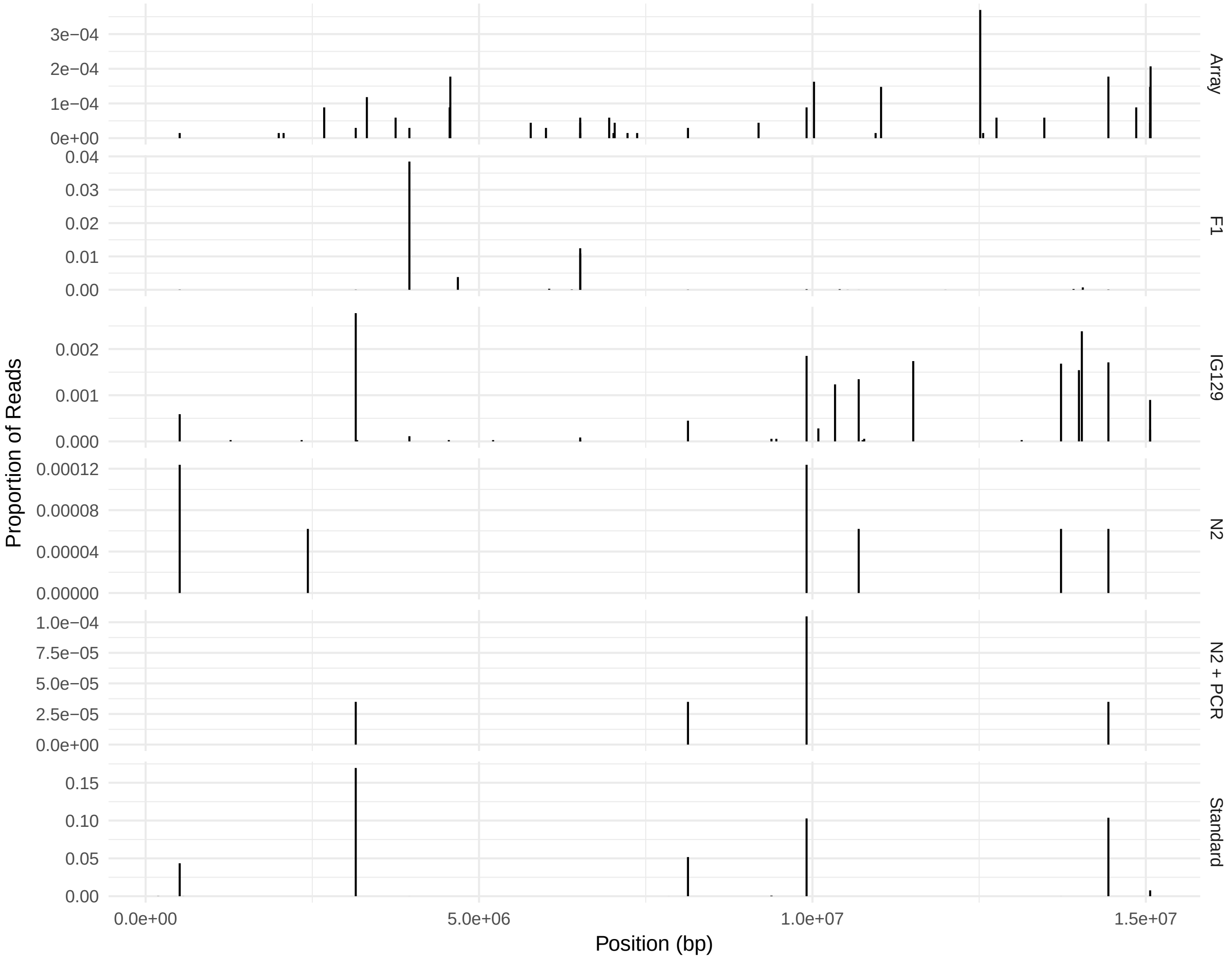
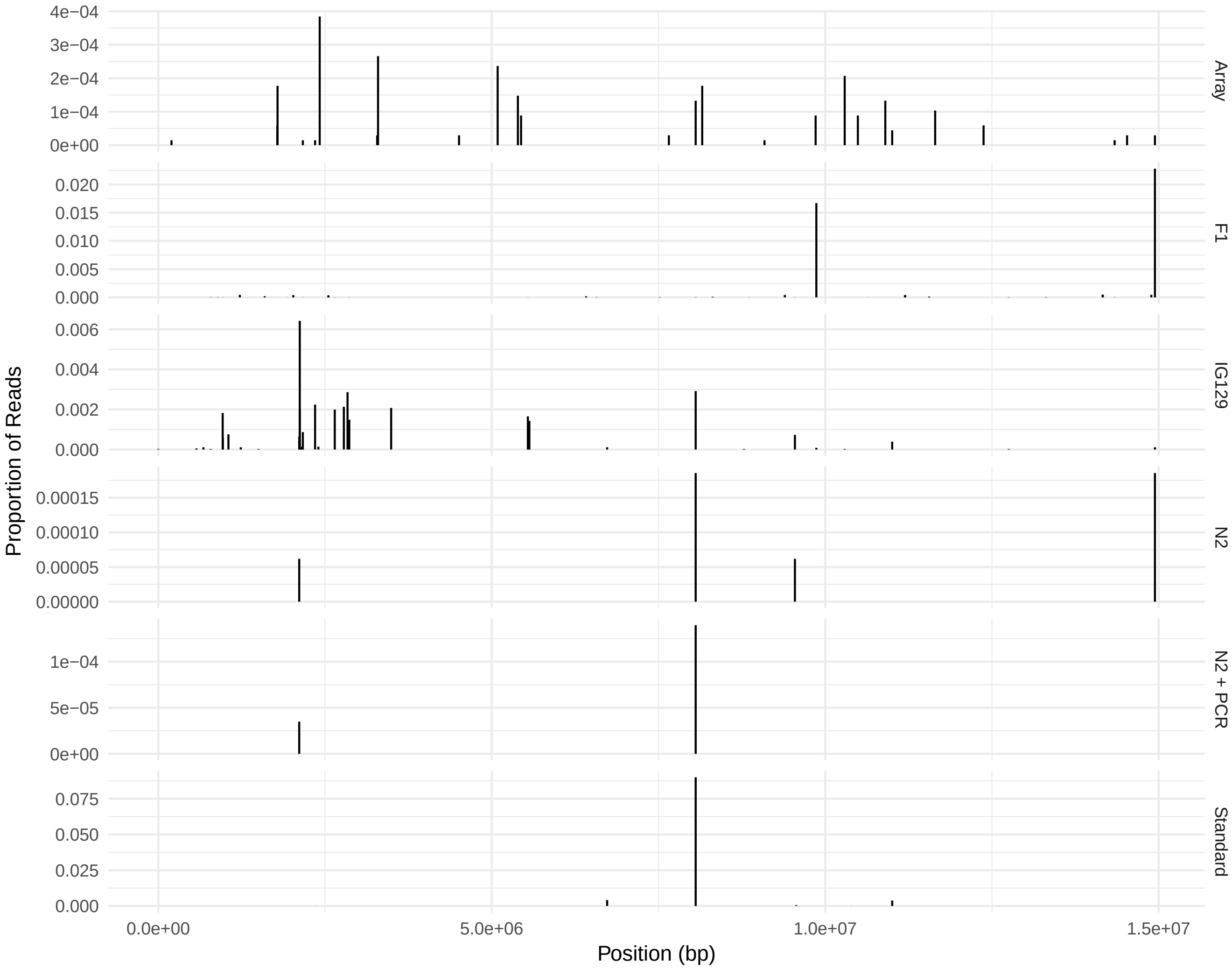


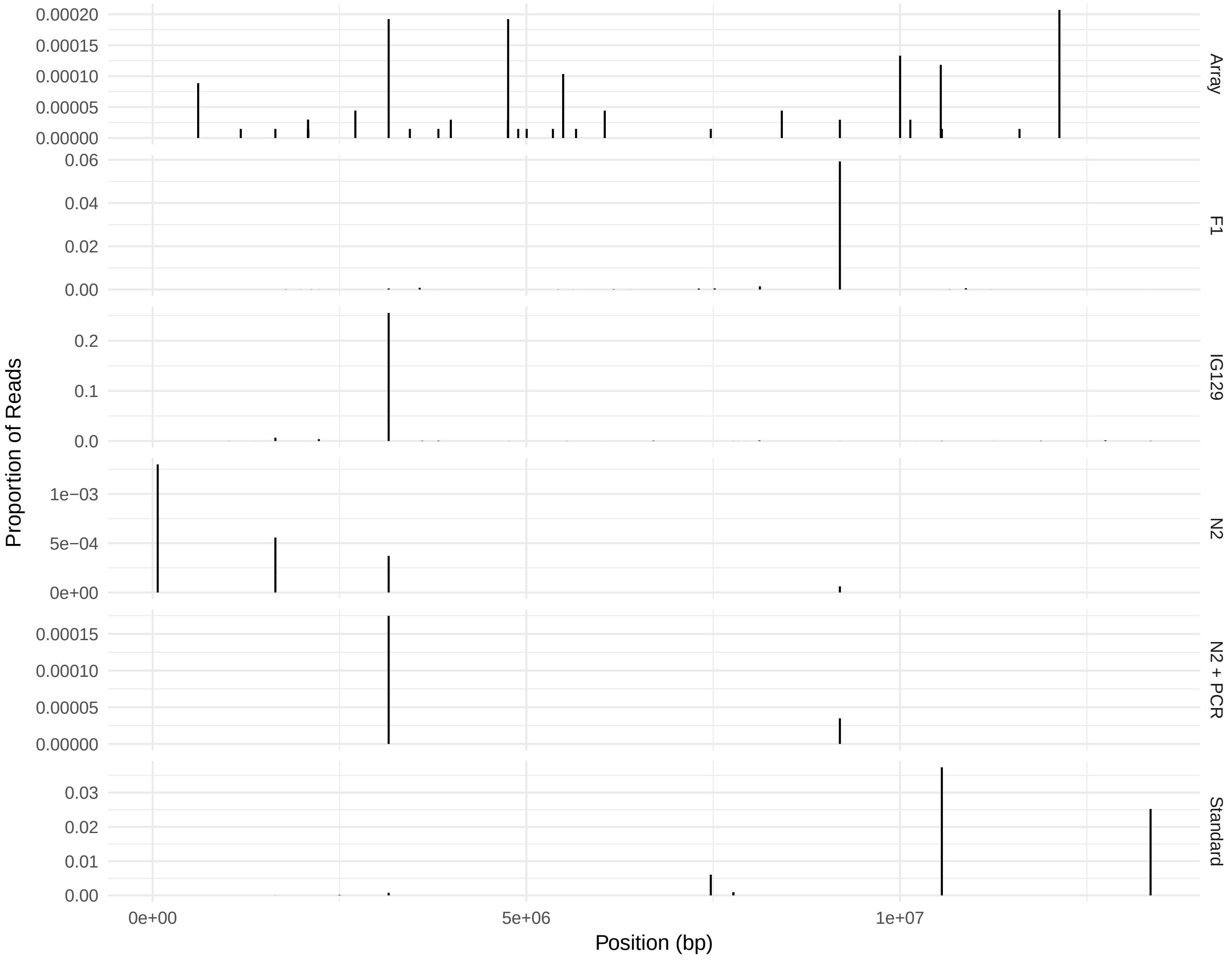
Figure 3.6: Control library gel (1.5% agarose). Lanes are, from left to right: N2 - N2 gDNA negative control, N2+ - N2 gDNA with transposon PCR product, Cb - combination of several clonal transposon strains, Sn - single clonal transposon strain (IG129), Arr - extra-chromosomal array containing strain (LRB311) with transposon at high copy number, F1 - a Ce-TnSeq mutant F1 population (100K animals), and L - NEB 100bp ladder. Note the faint band in lane N2 corresponding to cross-contamination of the transposon PCR product. The N2+ band corresponds to the expected size of the PCR product library fragment (407bp). Banding patterns were confirmed by Bioanalyzer (data not shown).



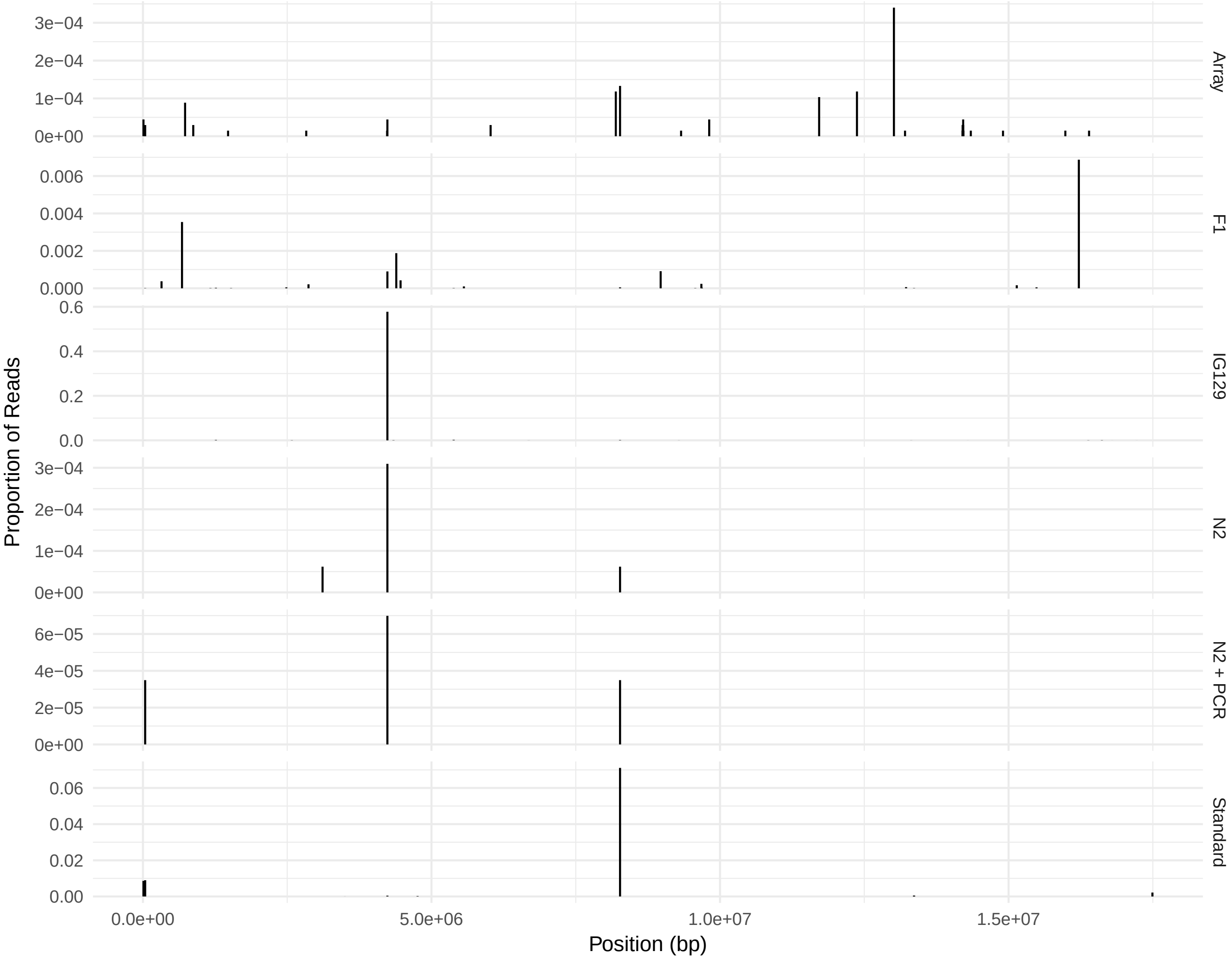
(a) Chromosome I



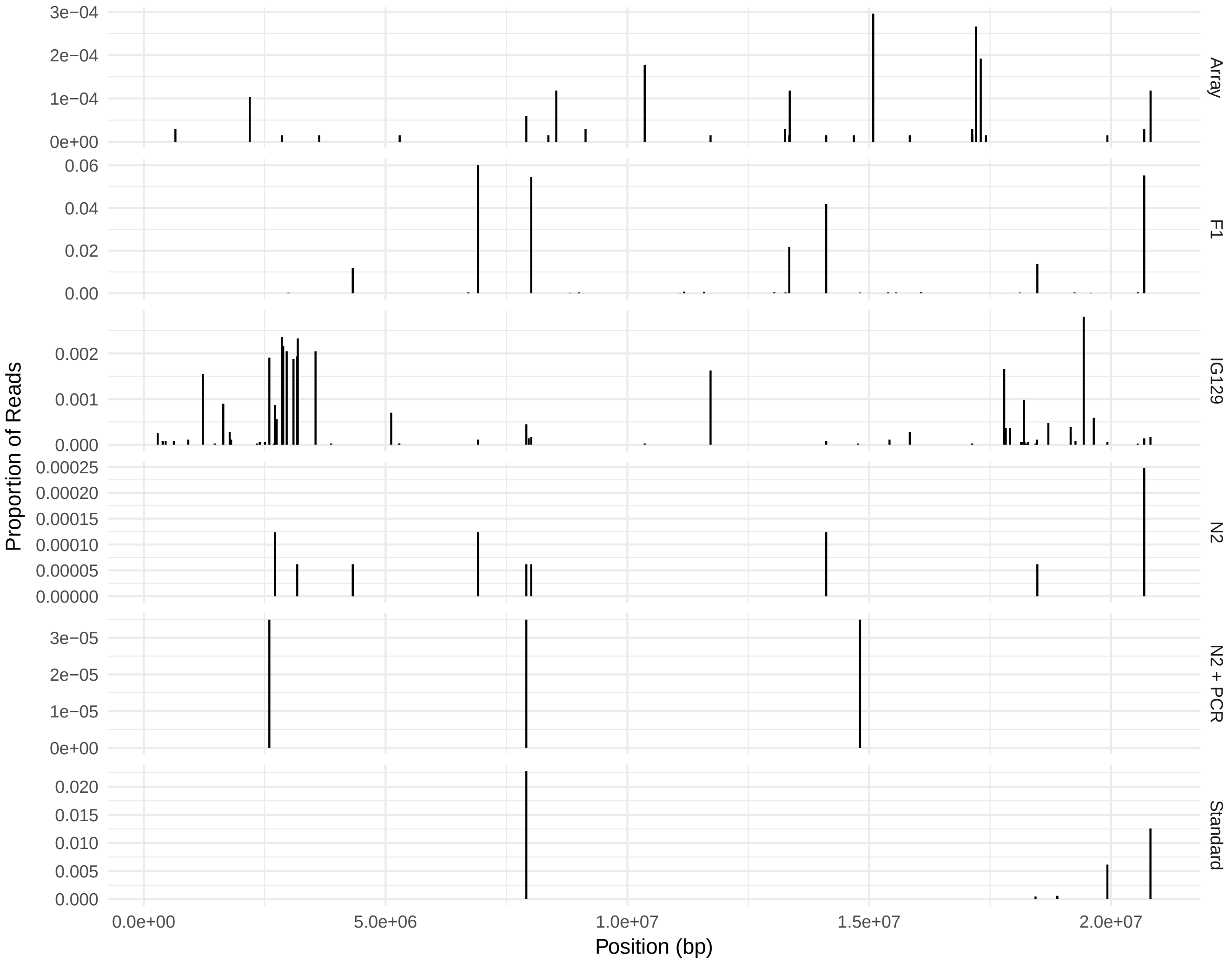
(b) Chromosome II



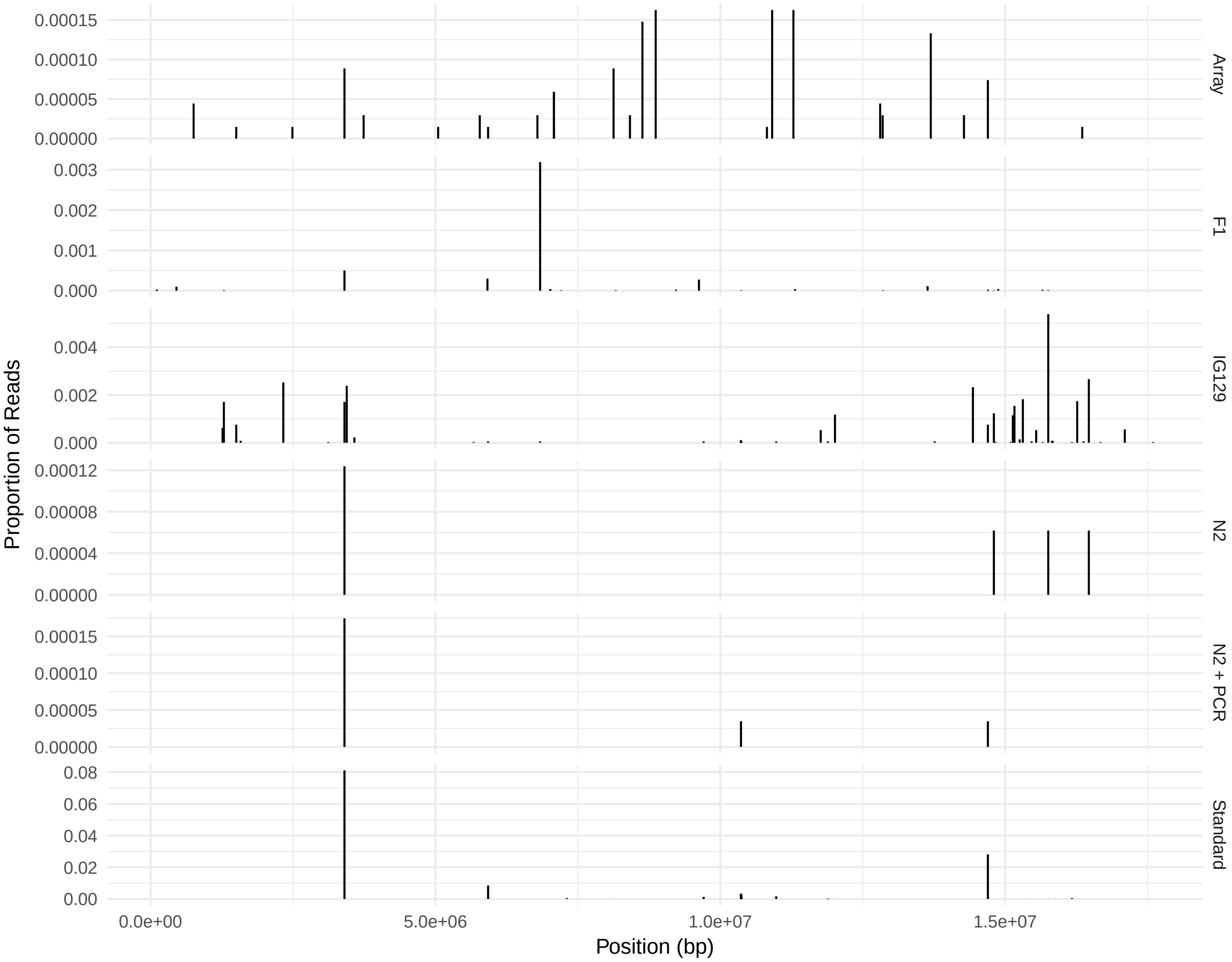
(c) Chromosome III



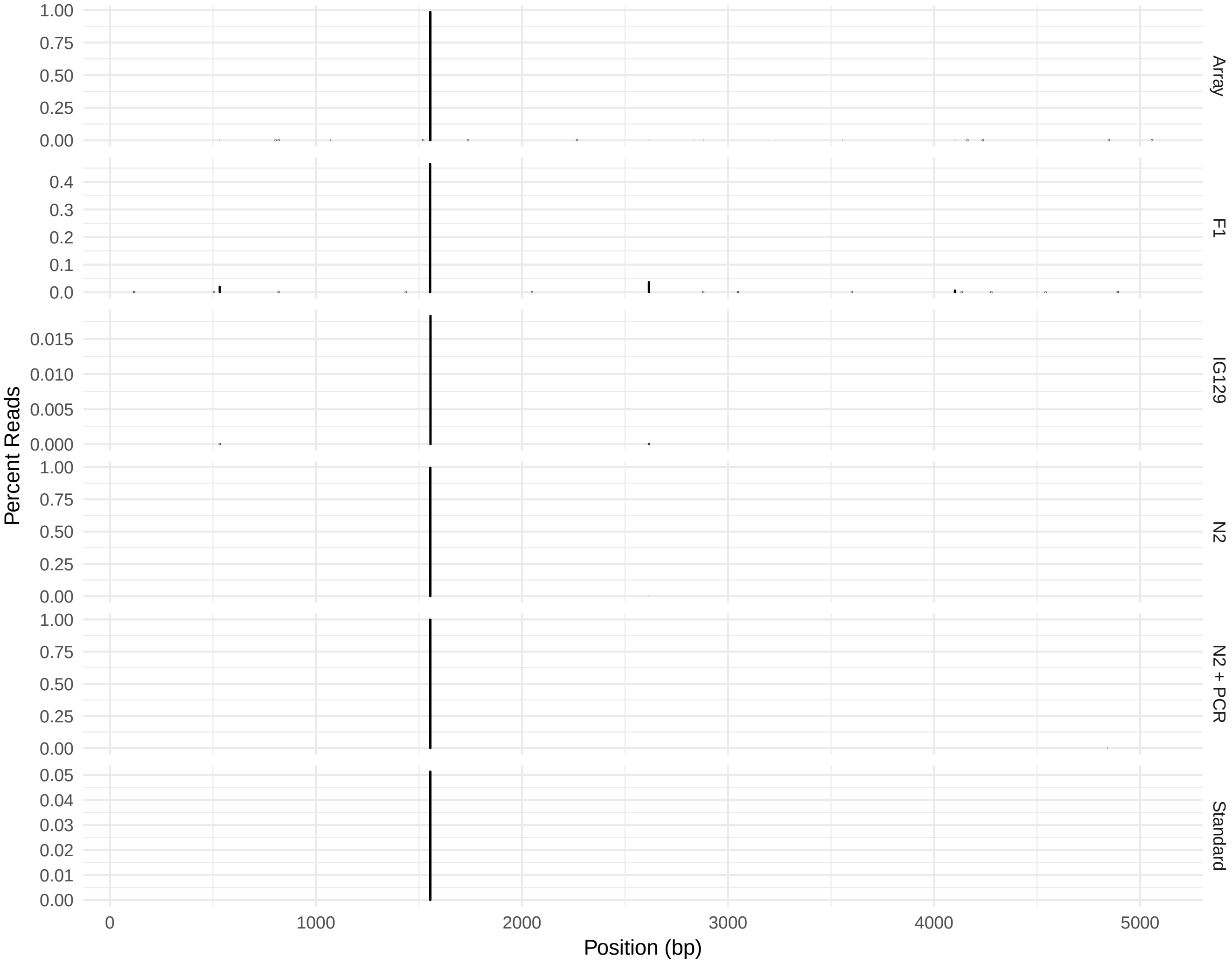
(d) Chromosome IV



(e) Chromosome V



(f) Chromosome X



(g) Plasmid pBM020

Figure 3.7: Location of reads in control libraries. Each subfigure denotes a separate chromosome. Note, the y-axis varies between samples and is the percentage of total reads coming from a particular locus (x-axis).

Another possibility is that Ce-TnSeq protocol failed to representatively amplify the insertion sites in the samples. That is, perhaps only a subset of the actual insertion sites in the F1 sample were detected in our library. Indeed, if we group our read pairs by the 5' end of read 2 (which should correspond to the location of random physical shearing of the original ligated molecule), we see the majority of reads sharing the same amplified molecule (on average, 100-fold amplification of each molecule). However, this explanation for a lack of diversity in the F1 is contradicted by our ability to detect the lowest concentration strains in the Standard sample as well as the fact there is no discernible difference in complexity (213 unique loci) from the other samples (Table 3.1).

Qualitatively, we were successful in identifying the individual strains in the Standard sample. During our analysis, we discovered the annotation quality of the strains to be poor- 2 of the 15 strains had reported insertion sites that would not BLAST to the current version of the worm genome, several of the other sites mapped to multiple locations in the genome. Despite this, we detected 14 out of the 19 (censoring the 2 previously mentioned strains) insertion sites reported for the Standard sample (some strains had more than one insertion site).

A requirement for any insertion site sequencing protocol is that the reads are correlated with the allele frequency in the sample. PCR amplification has known sequence-specific biases as well as a large amount of noise at higher cycles Kivioja et al. (2012)). Indeed, much of our library was comprised of PCR duplicate reads (Table 3.1). When we looked at the correlation between total reads at a locus and unique read 2 5' ends (unique reads), we saw low congruence (Figure 3.8). Further-

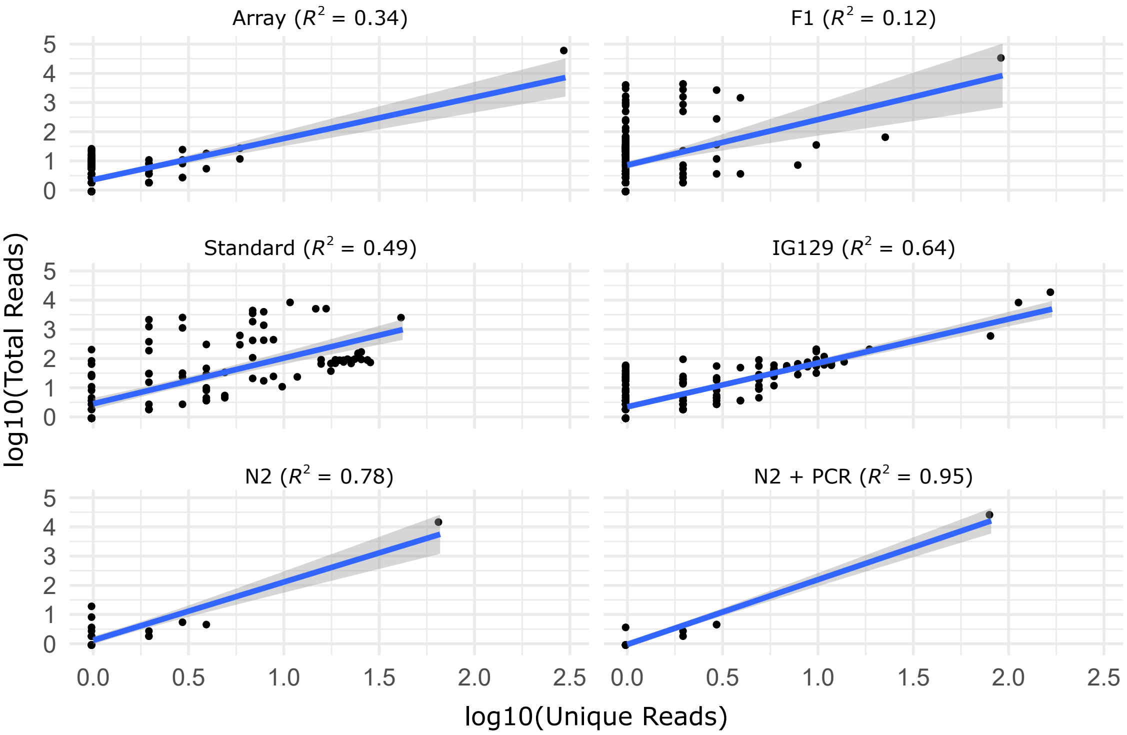


FIGURE 3.8: PCR duplicates identified in read data. For each of the control libraries, read 1 reads were grouped by location. The corresponding read 2 reads for these location groups were counted, and read 2 reads with the same 5' location were considered duplicates. X-axis is the log10 of the unique read 2 reads, and Y-axis is the log10 of the total read count (including duplicates). Each panel corresponds to a different control library.

more, when looked at the correlation between reads and the known concentrations of the strain in the Standard sample we also see low congruence ($R^2 = 0.38$, Figure 3.9). Restricting the read count to unique reads only marginally improves the correlation ($R^2 = 0.47$). This is probably explained by the low number of unique reads (2-18, Figure 3.9) per locus.

3.4 Resolving Duplicate Reads by a Universal Molecular Identifier

There are two ways we can improve the representativeness of the Ce-TnSeq protocol. First, we can optimize the number of PCR cycles. Indeed, we used an admittedly high number of cycles (12 PCR1, 30 PCR2) due to now resolved issues with the li-

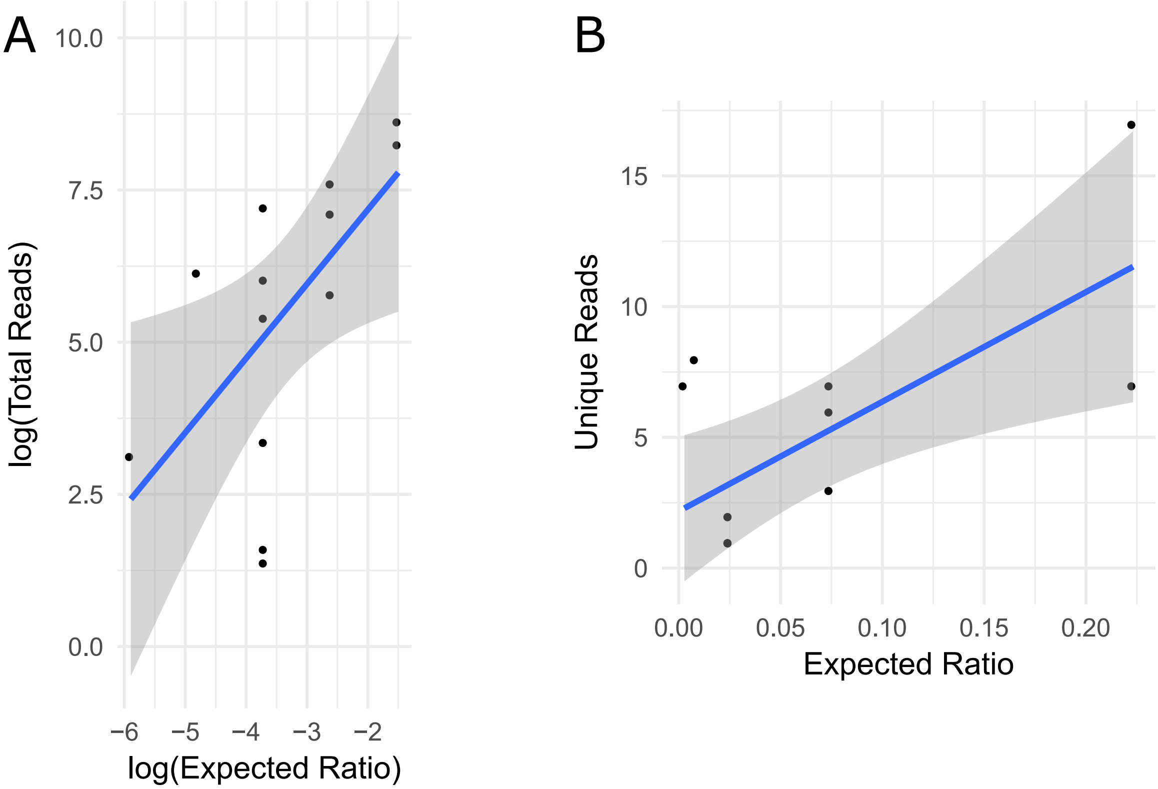


FIGURE 3.9: Inferring strain frequency by read count. Known concentrations of individual strains in the Standard sample (x-axis) are plotted against total reads per insertion site (a) or unique reads per site (b).

brary prep (incorrect sequence in one of the originally designed primers was causing the sequencer to fail). Therefore, we will use qPCR to identify the minimum number of cycles necessary to achieve sufficient amount of library DNA. Second, we have engineered adaptors with unique molecular identifiers (UMIs, also known as molecular barcodes, Kivioja et al. (2012)). A UMI is a random nucleotide sequence that is unique to each adaptor molecule. Each DNA fragment in the sample is marked with a different UMI at the adaptor-ligation step. These UMIs are conserved during PCR amplification, allowing the found molecule of each PCR duplicate to be traced. During sequencing, the number of unique UMIs at a particular locus are tracked instead of total read count. Assuming a low amount of sequence-specific bias at TA-ligation, these UMI counts should be more representative than the read counts.

UMIs have been used to increase the fidelity of mRNA counts in mRNA sequencing (Hashimshony et al. (2012); Ramsköld et al. (2012)). In CEL-seq (Hashimshony et al. (2012)) the poly-A primer used in reverse transcription contains a random barcode. In SMART-seq (Ramsköld et al. (2012)) a primer containing a random barcode is used in a template-switching reaction. In each case, the random barcode was generated during the synthesis of the corresponding single-stranded primer. The recent SiMSen-Seq protocol introduces UMIs by using barcoded primers in a 3-cycle PCR reaction prior to normal library PCR. Duplex sequencing (Schmitt et al. (2012)) uses an adaptor with a double-stranded random barcode. This barcoded adaptor is synthesized by annealing a short oligonucleotide with a homologous oligo that additionally contains a single-stranded barcode. The hybridized adaptor complex is made double-stranded by polymerase extension off the shorter oligo.

Here, we propose to introduce UMIs by synthesizing an adaptor molecule with a random *double-stranded* barcode, similar to duplex sequencing. We designed our UMI adaptor around the NEB DNA Ultra adaptor, in order to maintain compatibility with our existing protocol. We added a restriction enzyme target and random barcode to the 5' end of the NEB hairpin adaptor (Figure 3.10). The 3' end of the hairpin adaptor serves as our primer for polymerase extension. After second strand synthesis, we digested the adaptor with BciVI, a restriction enzyme that cuts 5 base pairs downstream of its target and leaves a single nucleotide overhang. This digest effectively leaves the hanging thymine necessary for TA ligation in the Ce-TNseq protocol.

Figure 3.11 shows the results of our synthesis. We saw the expected size increase after polymerase extension as well as the expected decrease after digestion. There was a an equimolar band of original template after polymerase extension, denoting a need to further optimize the reaction. We plan to optimize the reaction by increasing

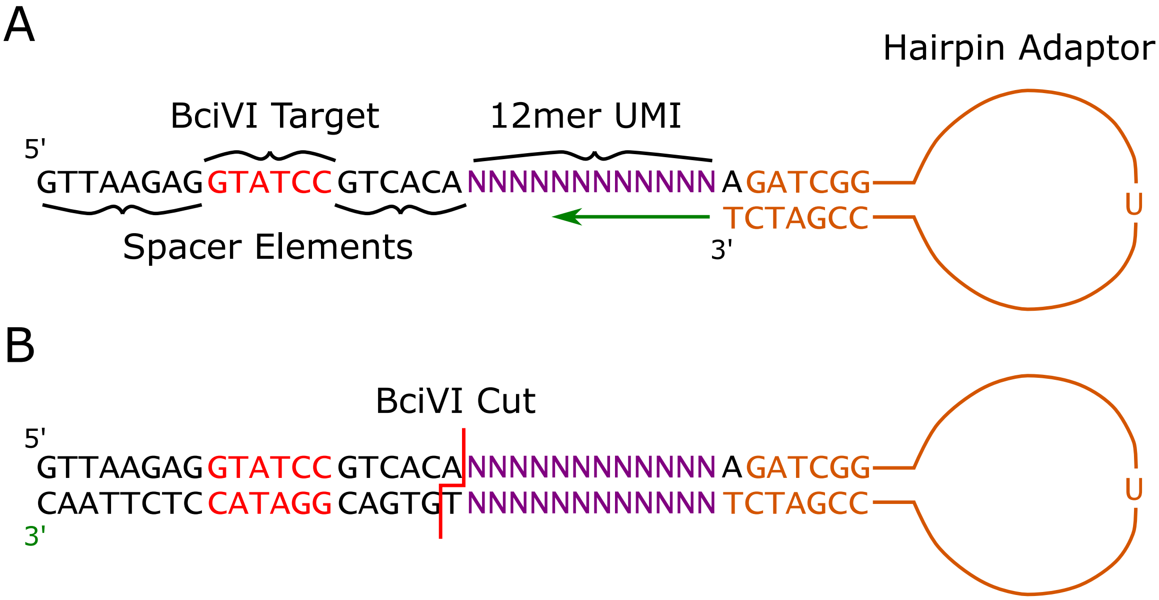

A
Hairpin Adaptor
BciVI Target
12mer UMI
5'
GTTAAGAG GTATCC GTCACA NNNNNNNNNNNN A GATCGG
TCTAGCC
U
3'
Spacer Elements
B
BciVI Cut
5'
GTTAAGAG GTATCC GTCACA NNNNNNNNNNNN A GATCGG
CAATTCTC CATAGG CAGTGT NNNNNNNNNNNN TCTAGCC
U
3'

FIGURE 3.10: Constructing a UMI sequencing adaptor. A. An existing hairpin adaptor is modified at its 5' end. The UMI (in purple) is designed adjacent to an off-cutting restriction site (BciVI, red). An isothermal polymerase (Klenow exo-) is uses the hairpin as a priming template and synthesizes the second strand (in direction of green arrow). B. Second strand synthesis is complete. A restriction digest with BciVI leaves a 3' thymine overhang on the adaptor and removes the extraneous sequence.

the amount of polymerase and increasing the extension time. If necessary, we can gel-purify the fully synthesized adaptor. Our UMI adaptor should be interchangeable with the NEB kit adaptor, allowing us to easily use it in future experiments.

3.5 Conclusions

I have designed and we have implemented a *C. elegans* transposon sequencing method (Ce-TnSeq), the first such method in a metazoan model. We have constructed a high-throughput Mos1 mutator strain (LRB311) that uses a drug-selectable marker for enrichment, and engineered AscI cut sites to alleviate the need for selection against array-containing mutants. This strain is necessary for our Ce-TnSeq protocol, but may be used as a more convenient alternative to the previous Mos1 strain (Boulin and Bessereau (2007)) for standard forward genetic screens.

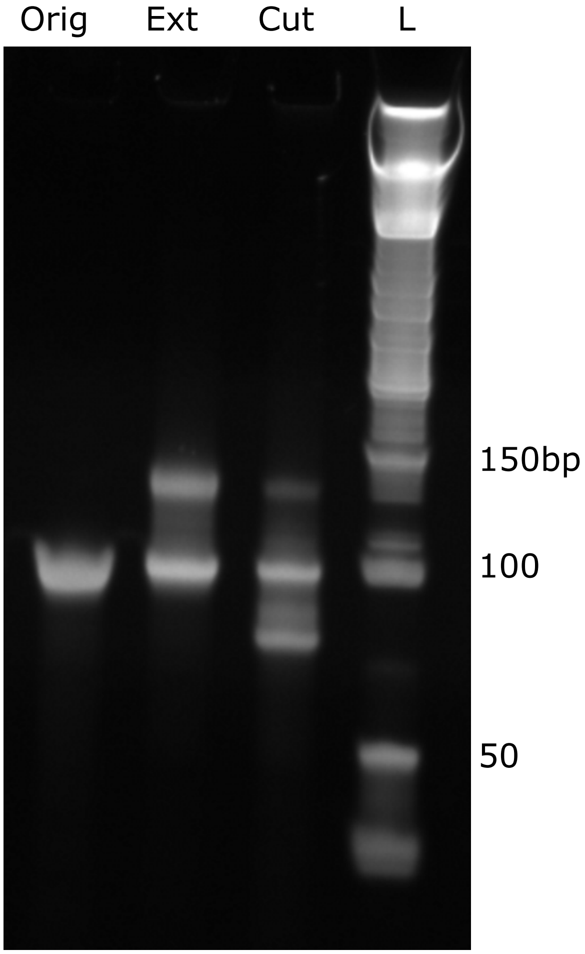


Figure 3.11: PAGE-Urea gel of UMI synthesis. Lanes are, from left to right: Orig - unmodified oBM109, Ext - adaptor after Klenow (exo-) extension, Cut - adaptor after extension and digest with BciVI, and L - 50bp ladder. Original adaptor is 97bp, after synthesis should be 130bp, after cutting should be 110bp.

Our Ce-TnSeq protocol is a modification of the commercial NEB DNA Ultra kit, introducing an restriction digest step (to remove the array) and replacing the standard PCR with a nested PCR specific to transposon insertion sites. We have shown that our protocol is qualitatively capable of identifying multiple insertion sites within a sample. We have identified the high amount of PCR amplification and the resulting large amount of PCR duplicates as an area for further optimization. We have discovered a surprising lack of insertion sites in an F1 control sample. We propose further replication of the mutation rate of LRB311 as well as Ce-TnSeq of additional F1 replicates.

Finally, we propose the introduction of UMI sequencing adaptors to improve the representativeness of our Ce-TnSeq reads. We have engineered and synthesize an UMI adaptor based off the commercial NEB sequencing adaptor. We have identified the polymerase extension step of the synthesis as a point for further optimization.

3.6 Materials and Methods

Table 3.2: Oligonucleotides and DNA sequences for Ce-TnSeq.

	Name	Description	Sequence
1	oBM070	Array PCR F	CGA CAT TTC ATA CTT GTA CAC CTG ATA
2	oBM098	Array PCR R	GAT AGG CAT CCC ACA GTA CGG
3	oBM095.F45	Tn PCR F	GCC GAA CTG CAA GCA TTA TT
4	oBM095.R180	Tn PCR R	CCC ATC TAC CGA CCT TCT GA
5	BM138	Sequencing Primer	TAT ATG TTC GAA CCG ACA TTC CCT ACT TGT ACA CCT GGT A
6	BM144	PCR1 Reverse Primer	GAC TGG AGT TCA GAC GTG TGC
7	BM146	PCR2 Forward Primer	AAT GAT ACG GCG ACC ACC GAG ATC TAC ACG TTT GCG AGA CAT CTA TAT GTT CGA ACC GAC ATT CCC
8	B7	PCR1 Forward Primer	TTT GCG TTT GAG CAT CGT CTT CAT C
9	NEB Adaptor		/5Phos/GAT CGG AAG AGC ACA CGT CTG AAC TCC AGT C/ideoxyU/A CAC TCT TTC CCT ACA CGA CGC TCT TCC GAT C*T
10	NEB Universal Primer		AAT GAT ACG GCG ACC ACC GAG ATC TAC ACT CTT TCC CTA CAC GAC GCT CTT CCG ATC*T
11	NEB Index 1	PCR2 Reverse Primer	CAA GCA GAA GAC GGC ATA CGA GAT CGT GAT GTG ACT GGA GTT CAG ACG TGT GCT CTT CCG ATC*T
13	Illumina Read 2 Primer		GTG ACT GGA GTT CAG ACG TGT GCT CTT CCG ATC T
14	Illumina Read Index Primer		GAT CGG AAG AGC ACA CGT CTG AAC TCC AGT CAC
15	Illumina Read 1 Primer		ACA CTC TTT CCC TAC ACG ACG CTC TTC CGA TCT
16	UMI Adaptor		/5Phos/GTT AAG AGG TAT CCG TCA CAN NNN NNN NNN NNA GAT CGG AAG AGC ACA CGT CTG AAC TCC AGT C/ideoxyU/A CAC TCT TTC CCT ACA CGA CGC TCT TCC GAT CT

3.6.1 Cloning pBM020

Genomic DNA from EG1470 (CGC) was amplified via PCR with Q5 polymerase (NEB, Cat. M0491S) and program 98 °C 1 m, [TD, 98 °C 15 s, 62-57 °C 15 s, 72 °C 40 s]x10, [98 °C 15s, 57 °C 15 s, 72 °C 40 s]x20, 72 °C 2 m using the following primer pairs: oBM103/oBM104, oBM105/oBM106, and oBM108/oBM109. The 3 products were added to gel-purified SmaI-cut pUC19 (Norrander et al. (1983)) vector at a 1:1:1:6 ratio to a total of 0.2 pmol DNA. Fragments were assembled and cloned using the NEB HiFi DNA Assembly Cloning Kit (NEB, Cat. E5520S) to create pBM020. See supplemental data for plasmid map.

3.6.2 Creation of strain high-throughput Mos1 mutator strain (LRB311)

Used standard procedure for microinjection (Mello and Fire (1995)). Injection mix: pCFJ90 Pmyo-2::mCherry::unc-54 UTR 1.7 ng/μL (Frøkjær-Jensen et al. (2012)), pJL44 Phsp16.48::MosTase::glh-2 UTR 68 ng/μL (Bessereau et al. (2001)), pBM020 AscI::Mos::AscI 35 ng/μL and pSG120 Prps-0::hygR 50 ng/μL (Tas et al. (2015)) to create LRB311 (dukEx125[Prps-0::hygR+Pmyo-2::mCherry+Phsp16.48::MosTase+Mos1 AscI Substrate]).

3.6.3 Hygromycin Selection

Gravid LRB178 culture was bleached and eggs were collected. Eggs were diluted to 10 μL in a 2X HB101 and 250 μg/mL hygromycin liquid culture. Culture was incubated with agitation at 20 °C until synchronized animals were L4/young adult. Culture was washed, and then plated at 1000 worms per pre-seeded OP50 10 cm plate. Plates were then ready for mutagenesis.

3.6.4 Mutagenesis

Plates of hygromycin-selected LRB311 young adults were subjected to heatshock at 33 °C for 1 hour in an air incubator. Plates were then moved to 20 °C for 1 hour, and then returned to 33 °C for an additional hour. Plates were recovered at 20 °C overnight. Plates were then washed (leaving behind eggs laid overnight) and bleached to collect synchronized F1 mutant progeny.

3.6.5 Mutation Rate

Adapted from (Boulin and Bessereau (2007)). We isolated 30 random F1 progeny from heat-shocked LRB311 parents. We allowed the individual progeny to self for 2 generations. From each F3, we picked 10-12 mCherry-negative animals into 10 μL of 1X ThermoPol PCR buffer (NEB) and 100 μg/mL proteaseK. We freeze-thawed the tubes in liquid nitrogen 3X, then incubate the tubes 65 °C for 1 hour, 95 °C for 30 minutes. We added 3 μL of each tube to PCR reactions with the following primer pairs (daf-16, BM095, BM070/BM098).

3.6.6 Genomic DNA Prep

We thawed each sample worm pellet (flash frozen in -80 °C freezer or liquid nitrogen, up to 300 μL). We added 180 μL of ATL buffer (Qiagen, Cat. 19076) and 20 μL of Proteinase K 20 mg/ml (NEB, Cat. P8107S) to the pellet and vortexed vigorously. We incubated in a heat block at 56 °C for 3 hours, vortexing occasionally. We added 4 μL of RNase A 100 mg/ml (Qiagen, Cat. 19101), mixed well, and incubatee at room temperature for 10 minutes. The solution was immediately transfered to a pre-spun Phase Lock Gel Heavy 2 mL tube (5 Prime, Cat. 2302830) and we added 750 μL of 25:24:1 buffered phenol:chloroform:isoamyl alcohol solution. We wrapped tube caps in Parafilm and vortexed vigorously. Tubes were spun at maximum RPM for 5 minutes. The top aqueous phase was transferred to a clean Phase Lock tube and the phenol:chloroform step repeated. We transfered the aqueous phase to a new Phase Lock tube and added 750 μL of chloroform. We vortexed vigorously and spin at max RPM for 5 minutes. We transferred the aqueous phase to a 1.5 mL microcentrifuge tube.

We added 5 M NaCl (note, using sodium acetate will cause an insoluble precipitate to form) up to 0.2 M and mixed well. We then added 2 volumes of 100% ethanol and mixed by inversion. We incubate the tube at -20 °C for 30 minutes. The tube was spun for 30 minutes at max RPM at room temperature. We removed the supernatant while carefully avoiding DNA pellet. We washed the pellet with 1 mL of fresh 70% ethanol and spun at max RPM for 1 minute. We removed the supernatant and repeated the wash. We air dried the pellet (careful not to overdry) and dissolved with 1 mM Tris-HCl pH 8 (aiming for 200-500 ng/μL final DNA concentration). To dissolve the pellet, we incubated it in a 40 °C water bath for 1 hour, vortexing vigorously throughout. DNA was stored at -20 °C.

3.6.7 Covaris Sonication

We sonicated 50 μL of 200 ng/μL DNA 1 mM Tris in a microTUBE (130 μL; Covaris, Cat. 520045) using a Covaris S220 (intensity: 5, duty cycle: 5%, cycles per burst: 200, treatment time: 35 s). Target fragment size was 500 bp.

3.6.8 Ce-TnSeq Library Preparation

We added 3 μL of End Repair Enzyme and 6.5 μL of End Repair Buffer (NEB, Cat. E7370S) to 55 μL of DNA (3 μg, 1 mM Tris). We then incubated: 30 m at 20 °C, 30 m at 65 °C, held at 4 °C. We then added 15 μL Blunt/TA Master Mix, 2.5 μL NEBNext Adaptor, 1 μL Ligation Enhancer (NEB, Cat. E7445S, E7350). We incubated 20 °C for 20 m. We then added 3 μL of USER enzyme (NEB, Cat. E7350) and incubated at 37 °C for 15 m.

We added 55uL SPRIselect beads (Beckman Coulter, Cat. B23317) and mixed thoroughly. We incubated at room temperature for 5 minutes. Tube was placed on magnet until beads separated (5-10 minutes). We transferred the supernatant to new microtube. We added 25uL SPRIselect beads and mixed thoroughly. We incubated for 5 minutes, and placed tube on magnet. Supernatant was then discarded. We added 200uL of fresh 80% ethanol and incubated for 30 seconds. We then removed the supernatant and repeated. We then removed the supernatant and allowed beads to air dry (beads had a moist matte look). We then added 17uL of 1mM Tris and mixed well. Incubate at room temperature for 5 minutes. Place beads on magnet. After separation (5 minutes) transfer 15 μL of supernatant to a new microtube. We then added 74 μL ddH20. We then added 10 μL 10X CutSmart buffer and 1 μL AscI 10 U/μL (NEB Cat. R0558S). Tube was incubated for 1 hour at 37 °C and then overnight at room temperature.

We added 80 μL SPRIselect beads (Beckman Coulter, Cat. B23317) and mixed thoroughly. We incubated at room temperature for 5 minutes. We places tube on magnet until the beads separated (5-10 minutes). We discarded the supernatant. We washed with 200 μL fresh 80% ethanol two times. We allowed beads to air dry and added 17 μL 1 mM Tris. We incubate at room temperature for 5 minutes. After separation, we transferred 15 μL of supernatant to a new microtube.

We combined 10 μL 5X Q5 reaction buffer, 0.5 μL Q5, 1 μL 10 mM dNTP, 1.25 μL B7 (20uM), 1.25 μL BM144 (20 μM), 21 μL ddH20. PCR program used was: 98 °C 1 m, [98 °C 15 s, 66 °C 15 s, 72 °C 45 s]x12 cycles, 72 °C 2 m. We added 40 μL SPRIselect beads. We incubated for 5 minutes at room temperature then placed tube on magnet. We discarded supernatant. We washed beads with 200 μL of fresh 80% ethanol two times. We allowed beads to air dry and added 17 μL 1 mM Tris. We incubated at room temperature for 5 minutes. After separation, we transferred 15 μL of supernatant to a new microtube.

We added 10 μL 5X Q5 reaction buffer, 0.5 μL Q5, 1 μL 10 mM dNTP, 1.25 μL BM146 (20 μM), 2.5 μL NEB Index Primer (10 μM), 19.75 μL ddH20. PCR program used: 98 °C 1 m, [98 °C 15 s, 72 °C 45 s]x30 cycles, 72 °C 2 m. We purified the PCR product with 40 μL SPRIselect beads as above. This was the library ready for further quality control and Illumina sequencing. BM138 was the custom sequencing primer for read 1.

3.6.9 UMI Adaptor Synthesis

First, we prepared 100 μM oBM109 by heating for 2 minutes at 90 °C, and then turned off the heat-block and let cool to room temp (which caused the hairpin structure to form correctly). For the extension, we combined 10 μL of 10 CutSmart Buffer (NEB, Cat. B7204S), 1 μL of 10 mM dNTP, 6 μL Klenow (exo-) polymerase (NEB, Cat. M0212S), 4 μL 100 μM oBM109 (PAGE-purified), 79 μL ddH20. Solution was incubated at 37 °C for 1 hour, and we heat-inactivated the polymerase at 75 °C for 20 minutes. We then added 2 μL of BciVI (NEB, Cat. R0596S). Finally, we incubated at 37 °C for 1 hour.

4

Population Sequencing of Wild Isolates to Study Natural Genetic Variation in Starvation Survival

The nematode, *Caenorhabditis elegans* is a popular and powerful model organism, but until recently the resources necessary to study the effect of natural genetic variation on on quantitative traits had not been developed for this animal (Gaertner and Phillips (2010); Andersen et al. (2012, 2015); Cook et al. (2016)). Recently, a collection of 97 wild isolates (Andersen et al. (2012)) was curated and genotyped by restriction-site associated DNA sequencing (RADseq). Though the genetic diversity of C. elegans was reportedly low, the authors were able to map alleles with large phenotypic effect (abamectin-resistance and aversion to the pathogen *Pseudomonas aeruginosa*) through association studies. These association studies were labor-intensive, requiring the maintenance and phenotypic screening of the 97 strains individually and with replication.

We sought to apply the ideas of transposon sequencing, selection and screening on entire populations as well as population sequencing to measure genetic shifts, to increase the throughput of association studies in C. elegans. Instead of sequencing transposon insertion sites, we measured SNP frequencies at restrict site-associated loci. Furthermore, instead of measuring the change in *allele* frequencies due to selection, we sought to measure the shift in *strain* frequencies within the population. We would then use the shift in strain frequency as our quantitative trait in our association study.

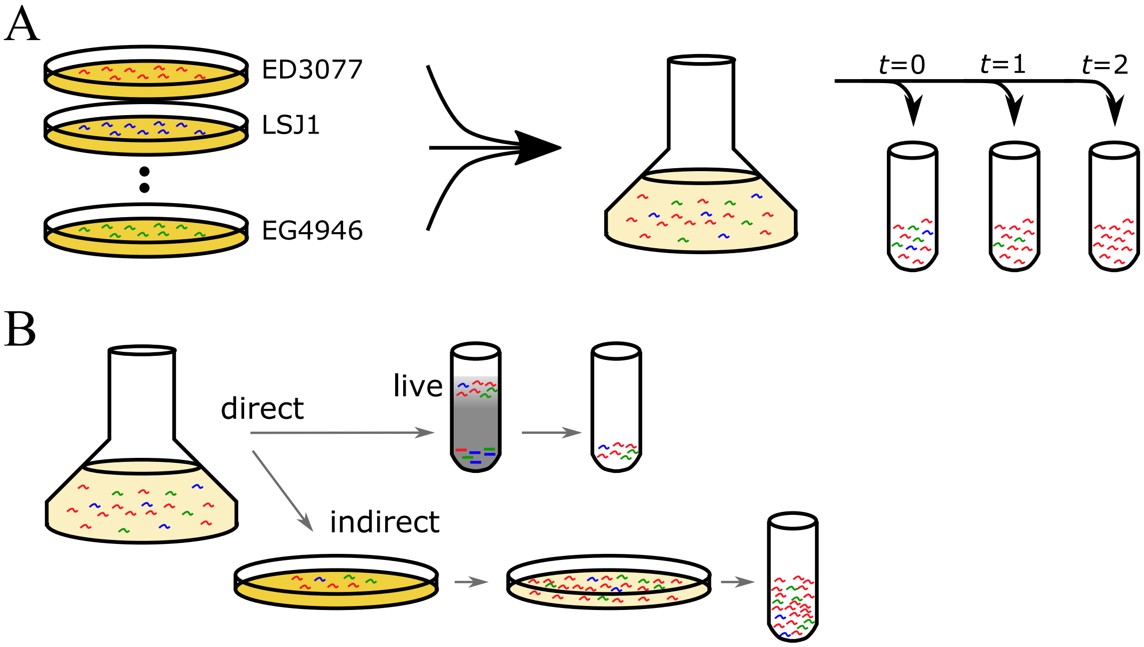


FIGURE 4.1: Population sequencing experiment design. A. 95 wild-isolate strains were grown individually on plates (e.g. strains ED3077, LSJ1, and EG4946). Prior to the experiment, each plate is washed into a single pooled liquid culture. The culture is bleached-synchronized, and then samples are taking at different time points throughout starvation. B. Samples from the starved culture are subject to either direct or indirect scoring. For direct scoring, sucrose flotation is used to separate live and dead worms based on density. The living worm portion is then sequenced. In indirect scoring, an aliquot containing live and dead worms is placed onto a plate with food. The worms are allowed to grow until the food is exhausted (1-2 generations of growth). The starved plate is washed and then sequenced.

4.1 An Association Study on Starvation Survival using Population RADseq

An overview of our experiment design is show in Figure 4.1. We maintained isolated cultures for 95 of the 97 strains reported in Andersen et al. (2012). Prior to our screen for starvation survival, we washed each strain plate into a combined liquid culture. This population of pooled strains was propagated for 2 generations and then synchronized by bleaching. The resulting synchronized culture of eggs were allowed to hatch in the absence of food causing them to arrest at the L1 larval stage.

The starved culture was incubated at 20 °C under agitation, and samples were taken every 7 (replicate 2) or 8 (replicate 1) days. Figure 4.1B depicts the handling of samples. At each time point, samples were subject to either *direct* or *indirect* screening for starvation survival, and the subpopulations of animals that passed screening were frozen for future population RADseq. For *direct* scoring, the sample containing live and dead worms were separated by density centrifugation (i.e. sucrose flotation, see Materials and Methods). The supernatant containing mostly live worms was reserved for sequencing. For *indirect* scoring, the sample containing live and dead worms was placed on a plate with food. The plated culture was allowed to grow until food was exhausted, then the entire plate was washed and reserved for sequencing. The indirect scoring was designed to capture an aspect of starvation survival more relevant to fitness, that is, the ability to recover from starvation and reproduce. Note, this is a complex phenotype as it is affected by the animal's growth rate and brood size after starvation.

During the first replicate, we observed that virtually no dead worms were found in the sucrose supernatant and that our efficiency of recovering live worms was 80-90% per time point. However, the sucrose flotation failed to work during the second replicate, with large proportions of live and dead worm intermixed in the supernatant and lower portion of the tube. One difference we observed between replicates was the presence crystals in the buffer in replicate 1 but not in replicate 2. Regardless, we censored the direct scoring samples from replicate 2. The experimental time points and replicates acquired are shown in Table 4.1.

Table 4.1: Replicates and time points.

Scoring	Replicate	Time Points (days)
Direct	1	1,8,12,16
Indirect	1	16,20
Indirect	2	1,7,14,21,28

4.2 Population RADseq

Restriction site-associated sequencing (RADseq) allows for a predetermined subset (defined by the restriction enzyme used) to be sequenced by next gen sequencing. By reducing the amount of DNA sequenced from the whole genome to the restrict site-associated loci, the effective coverage per animal is increased greatly (from the 100mb worm genome to the 45936 EcoRI sites). Andersen et al. (2012) previously identified 41K single nucleotide polymorphisms (SNPs) among the 97 isolates. Within those 41K SNPs, we identified 12285 SNPs that were unique to a single strain of the 97. The distribution of the number of unique SNPs per strain is shown in Figure 4.2.

In order to infer the strain frequencies within a population, we used the following method. We made a RADseq library from the population DNA. We aligned the reads from the RADseq library to the 12K unique SNP locations we identified. For a given strain, our measure of strain frequency is the median ratio of minor/major alleles of its unique SNP loci.

We sequenced two control libraries of $3 \times 6 = 18$ strains mixed at pre-defined proportions. In one library, strains were split amongst 3 groups (triplicate) and within groups were added in concentrations of a 1.5 fold standard (Figure 4.3). The other library was a 2-fold standard (Figure 4.4). We observe many cases where the median unique SNP ratio agrees with the standard concentration, but we also observe cases where the median ratio is offset from the expected concentration (bias) or had a long tail in the distribution (noise). We hypothesize that the bias in detecting

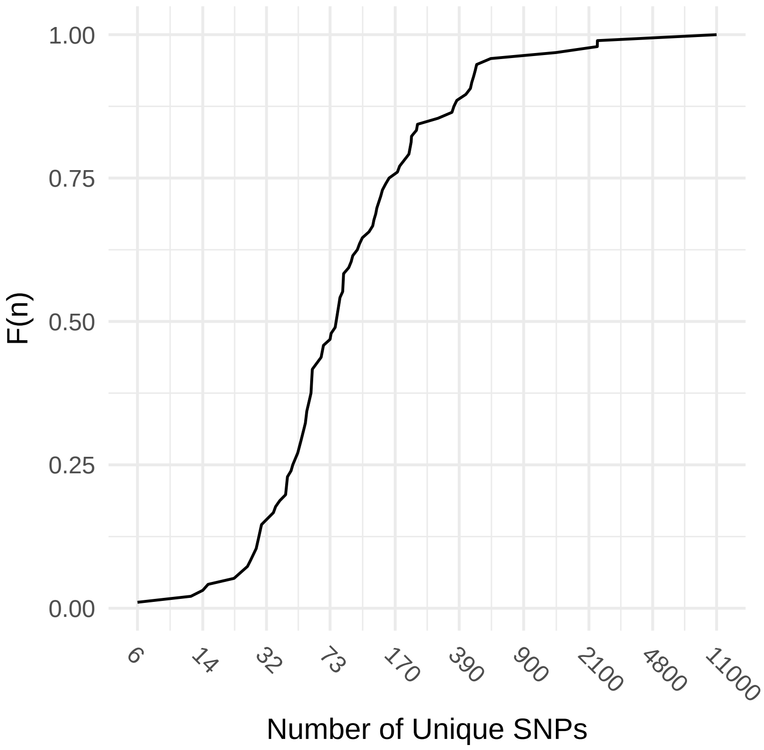


Figure 4.2: Cumulative distribution of unique SNPs. The distribution of the 28449 unique SNPs by strain is shown. For example, 50% of strains have 73 or fewer SNPs that are unique amongst the 95 strains screened. The x-axis is on a log10 scale.

certain strains' frequencies could have been due to bacterial contamination in the DNA preps. For the standard, each strain's DNA sample was quantified using a DNA-binding fluorescent dye and spectrometer (Qubit). This quantification method cannot tell the difference between different sources of DNA. To resolve this in the future, we propose to use qPCR to quantify the standard's DNA. We hypothesize that the noise in unique SNP ratios could be due to PCR amplification noise.

We proceeded with sequencing the starvation survival samples in Table 4.1. The results are shown in Figure 4.5. We saw many examples of discordance between the direct and indirect scoring. This isn't surprising as the indirect scoring is confounded by growth rate and fecundity. The comparison between replicates is also problematic as we only had two time points near the end of the series. We did observe that

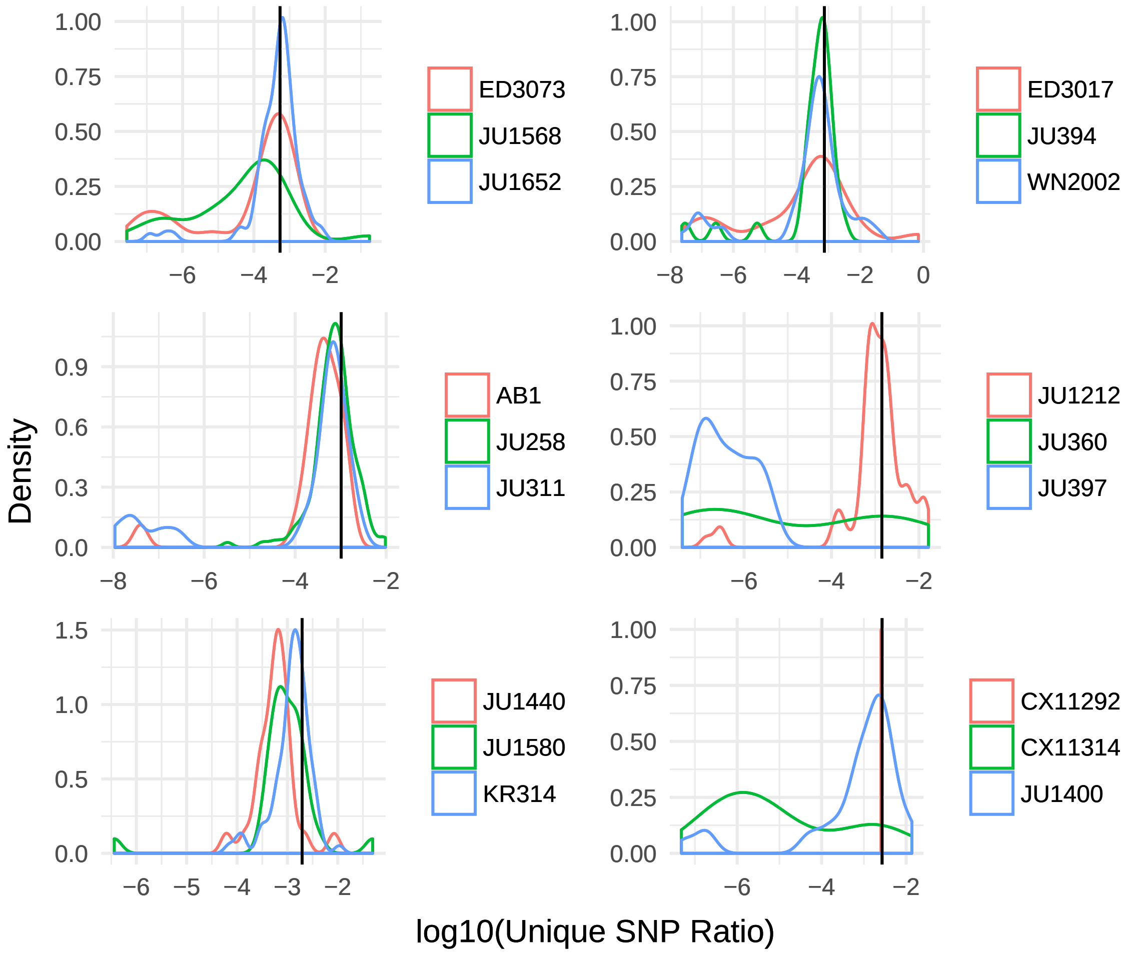


Figure 4.3: Distribution of 1.5-fold spike-in controls. Each panel is a different concentration of the 1.5 fold standard. Color differentiates the triplicate strains at each concentration. Black vertical lines denote the expected SNP ratio given the standard concentration.

ED3077 was a clear outlier in the indirect scoring. Despite starting at relatively low concentrations in the population, it dominated the later time points.

We decided to proceed with our association study. As we had the most time points in the indirect scoring of replicate 2, and had observed discordance with the other replicates and scoring methods, we limited our analysis to the replicate 2 data. We fit a line to the time course data and used the slope as our quantitative trait (trend). Figure 4.6 shows the distribution of trend values across the strains. The

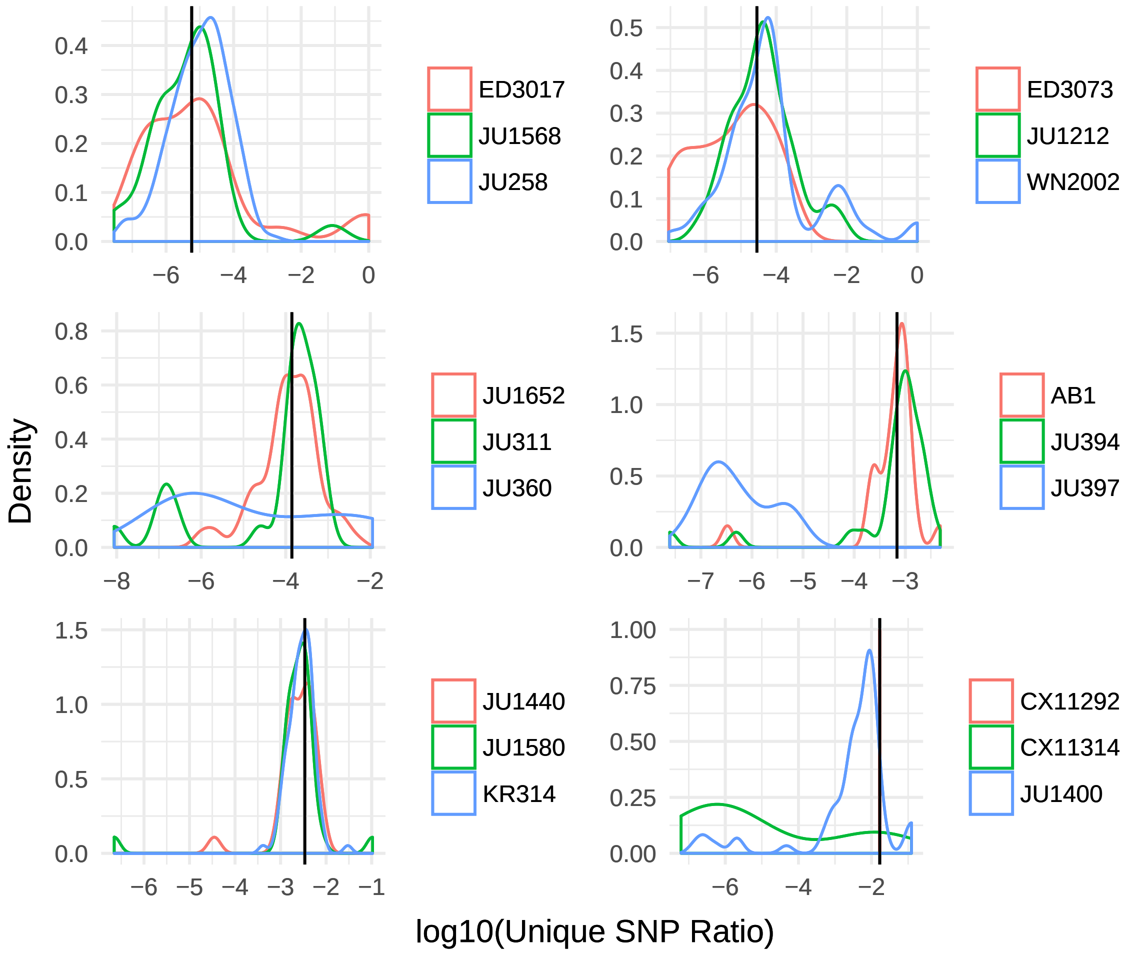


FIGURE 4.4: Distribution of 2-fold spike-in controls. Each panel is a different concentration of the 2 fold standard. Color differentiates the triplicate strains at each concentration. Black vertical lines denote the expected SNP ratio given the standard concentration.

rightmost outlier represents ED3077.

The *C. elegans* Natural Diversity Resource (CeNDR, Cook et al. (2016)) provides a tool for performing genetic mapping of quantitative traits on a growing set of natural isolates (including the 95 strains used in our study). It uses deep sequencing data for each strain to finely map quantitative trait loci (QTL). The results of the CeNDR analysis of our trend data is shown in Figure 4.7. We identified two significant peaks: one on chromosome I and one on chromosome III.

4.3 Validating the RADseq Results

We wanted to validate the results of the starvation survival assay. We 3 strains with relative high trend values and 2 strains with relatively low trend values. We then screened the strains individually for their response to starvation. Like the previous experiment, arrested L1 larvae were kept in buffer without food. At each time point, a 100 of the starved L1 larvae were plated with food. Worms that exited the L1 larval stage after 48 hours of exposure to food were considered "recovered". The results are shown in Figure 4.8. We found that the high trend strains outperformed the low trend strains, with 2 of the 3 high trend strains (ED3077 and DL200) showing statistically significant survival times. Note, ED3077 was an extreme outlier in our RADseq assay, but showed marginal improvement (2 days). This may be due to the difference in scoring methodology be the indirect scoring and the 48 hour recovery scoring. The 48 hour recovery period is too short for the animals to be reproductive, so brood size is not considered.

4.4 Conclusions

Using population RADseq, we identified strains with starvation survival phenotypes from a population of 95 strains. We measured strain frequency in the population by measure major/minor allele counts at SNP loci unique to each strain. We validated our approach by measuring survival in the identified outlier strains. We also performed an association study on starvation survival using survival trends inferred from our population RADseq data. Finally, we identified two significant QTLs for starvation survival.

4.5 Materials and Methods

4.5.1 Strain Maintenance

Prior to the starvation survival assay 97 strains were maintained on NGM agar plates seeded with OP50 *E. coli* and incubated at 20 °C.

4.5.2 Survival Assay

Isolated strains (maintained as above) were washed from plates and combined into 2X HB101/S-Complete media at 5 animals/uL. This culture was maintained for two generations, and then subjected to bleach treatment. Recovered eggs were resuspended in S-virgin buffer at a concentration of 10 eggs/uL. This culture was incubated at 20 °C with agitation. The animals were given 24 hours to hatch an arrest in L1, and then the first starvation survival sample was taken.

4.5.3 Direct Scoring

50uL aliquots were taken from the starved liquid culture and scored for live/dead animals by movement. The percentage of animals alive was used to estimate volume of aliquot should be taken to acquire 100K living worms. The aliquot was spun at 4000rpm for 1 minute and resuspended in 1mL S-virgin. The 1mL resuspension was then added gently to a 15mL conical tube containing 10mL of 30% sucrose. Without mixing, the solution was spun at 4000rpm for 5 minutes. With a large bore pipette, the top 2mL of supernatant was removed. The supernatant was washed 3X with 10mL of S-virgin. After the final wash, the solution was spun at 4000rpm for 1 minute, the supernatant removed, and the resulting worm pellet was frozen at -80 °C prior to sequencing.

4.5.4 Indirect Scoring

Percent of alive worms was calculated as above. A volume corresponding to 500 living worms was added to a seeded OP50 (with food) 10cm plate (10 plates per time point). Plates were incubated at 20 °C until the food was exhausted, 68 hours for the first time point, 96 hours for days 7 and above, and 170 hours for day 28. After food depletion, plates were washed and combined, then further washed 2X with S-virgin. After the final wash, the worm pellete was retrieved and frozen and -80 °C.

4.5.5 gDNA Isolation and Purification

Purified gDNA was retrieved from each worm pellet using the DNeasy Blood and Tissue kit (Qiagen, Cat. 69504) per manufacturer's instructions.

4.5.6 RADseq

Samples were prepared for RADseq as in Baird et al. (2008). Prepared samples were 25bp single-end sequenced on Illumina HiSeq 2500 V3 Rapid Run.

4.5.7 Processing and Analysis of RADseq data

Unique SNP frequencies were calculated as follows. Illumina read data was aligned to worm genome WS220 using bwa with default parameters (Li and Durbin (2009)). SNP calls were made using the samtools mpileup command (Li et al. (2011b)). SNP analysis was constrained to 12K unique SNPs from Andersen et al. (2012)).

4.5.8 48 Hour Starvation Recovery Assay

Individual strains were grown on separate plate on NGM+OP50 at 20 °C. Prior to the assay strains were synchronized using bleach treatment. Eggs collected were allowed to arrest at the L1 larval stage for 24 hours in S-virgin. Aliquots of 100 animals were taken at each time point and recovered on plates with food. Animals significantly larger than L1 larva were scored 48 hours after recovery.

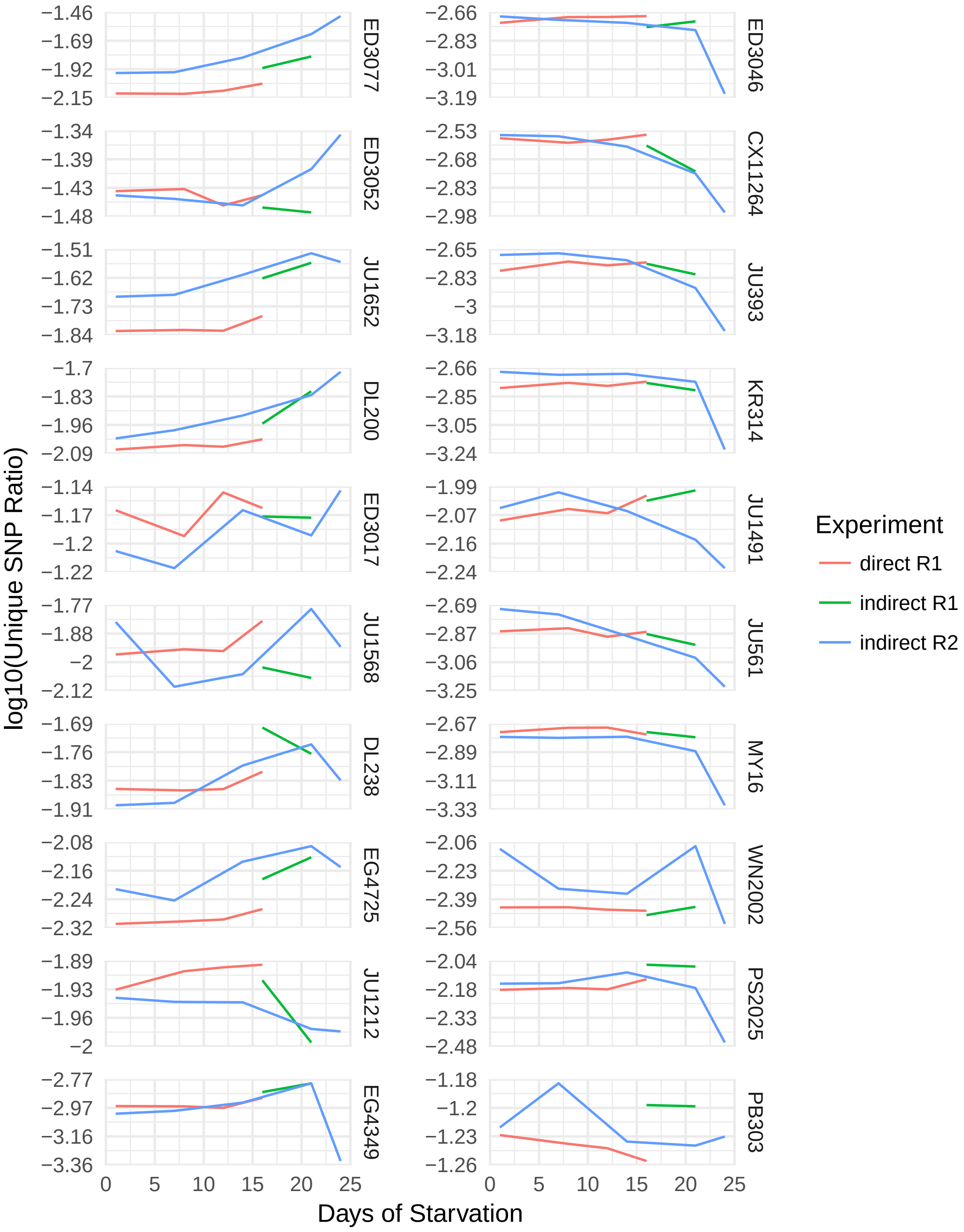


FIGURE 4.5: Strain frequency over starvation. The estimated strain frequencies from RADseq reads for the different replicates and scoring methods. Only a subset of the 95 strains is shown. On the left, the top 10 highest trend strains and on the right the lowest strains by trend. Trend was calculated as the slope of a linear regression fit to the replicate 2 indirect scoring time course.)

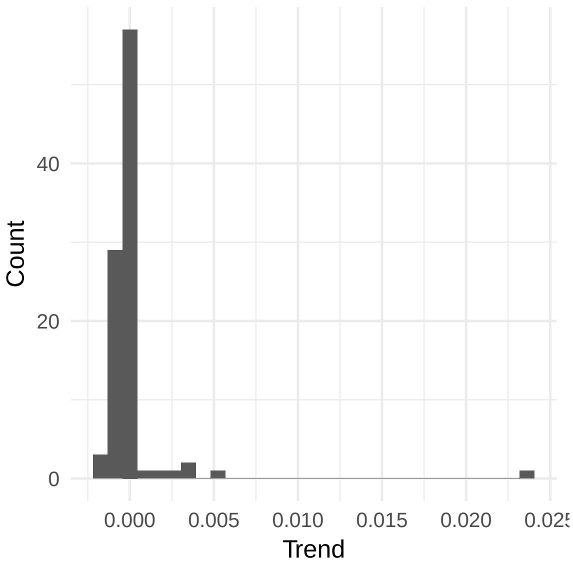


FIGURE 4.6: Histogram of trend values. Trend was calculated as the slope of a linear regression fit to the replicate 2 indirect scoring time course. The outlier on the right is ED3077.

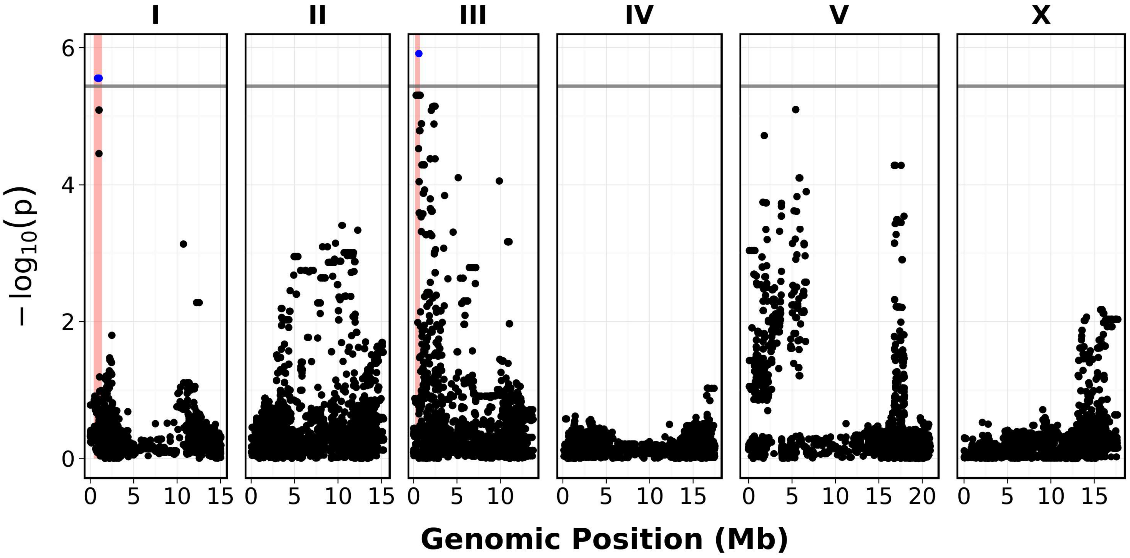


FIGURE 4.7: QTL analysis of trend data. Shown is output from the CENDR (Cook et al. (2016)) tool. Significant peaks are shown in red.

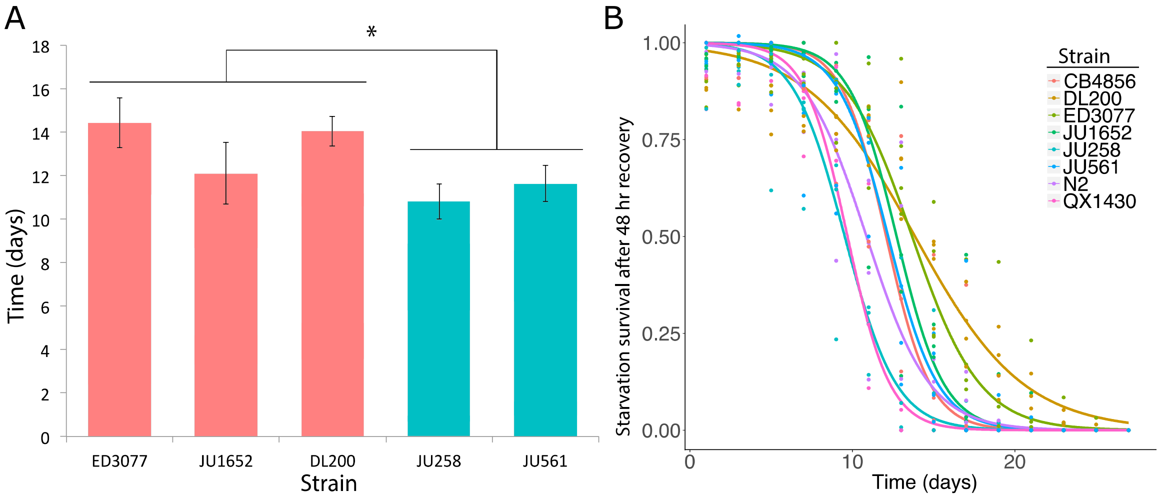


Figure 4.8: Survival data on outlier strains. A. Bar plot showing mean survival time for high RADseq trend strains (red) and low trend strains (blue). B. Survival curves for the various strains. Individual dots represent survival rates from different replicates and the lines are the data fit to a logistic curve. From Anthony Hung, unpublished.

5

Conclusions

5.1 Increasing the transposon mutation rate

In Chapter 2, I demonstrated that the mutation rate of the Mos1 transposon is quite low. On average, it causes insertions in just over 50% of animals, with an expected 1-2 mutations per mutant animal (Boulin and Bessereau (2007), Vallin et al. (2012)). I showed that this limited mutation rate leads to over 400K F1 animals being required to saturate the genome (Figure 2.1). Further results from the model showed that 50K animals was the upper limit on the number of F1 samples in a single experiment. The conclusion was that it would be necessary to replicate a TnSeq experiment eight times in order to saturate the genome. An eight-fold increase in mutation rate would be sufficient to saturate the genome in a single experiment.

The current Mos1 mutagenesis protocol was optimized with respect to the duration of heat-shock (induction of the transposase) and the window of egg-collection for F1 progeny (Boulin and Bessereau (2007); Williams et al. (2005)) in order to maximize the number of mutant progeny per P0 parent. It was concluded that increasing the heat-shock led to increased somatic mutation and sterility (Boulin and Bessereau

(2007)), while the egg-collection window was capturing embryos whose oocytes were meiotic in the germline pachytene at time of mutagenesis (Williams et al. (2005)). It is an open question whether the *in vivo C. elegans* Mos1 mutation rate can be further increased.

A primary concern in attempting to increase the Mos1 mutation rate is the claim that the sickness and infertility resulting from prolonged heat-shock is due to the increased rate of somatic mutation (Boulin and Bessereau (2007)). An alternative hypothesis is that the extended heat-shock itself caused infertility. If the sickness is indeed caused by somatic mutation, one solution would be to replace the ubiquitously expressed *hsp-16-48* heat-shock promoter (Stringham et al. (1992); van Luenen et al. (1993)) with a germline or pachytene-specific promoter and 3' UTR (e.g. *Ppie-1::gld-1 3' UTR*; Merritt et al. (2008)). Such a germline-specific system can be made inducible using FLP-Out (Voutev and Hubbard (2008)). Alternatively, if heat-shock itself is causing sickness, a heat-shock -less inducible system, such as cGAL (Wang et al. (2016)) could be used.

Several studies have been conducted on optimizing the rate of Mos1 transposition *in vitro*. It was shown that the Mos1 transposition rate is dependent on several internal sequences (not just the ITRs), and that most perturbations to the sequence or spacing of these internal elements resulted in a dramatic drop in mutation rate (Lohe and Hartl (2002)). Note, this result conflicted with previous work that had incorrectly used a modified Mos1 transposon (it had similar size to the wild-type Mos1, but contained an artificial insert) as a control for transposition rate (Tosi and Beverley (2000)). It has been shown that the *in vitro* rate is affected by temperature, with 28 °C being an optimal temperature and an order of magnitude more efficient than 25 °C or 30 °C (Sinzelle et al. (2008)). The wild-type ITRs of Mos1 are asymmetric, with the 5' ITR differing from the 3' by four bases (Augé-Gouillou et al. (2001)). It was shown that the *in vitro* efficiency of transposition could be increased

1000-fold by replacing the 5' ITR with the optimal 3' ITR sequence (TCA GGT GTA CAA GTA TGA AAT GTC GTT TCC C; Augé-Gouillou et al. (2001)). It is also known that germline silencing of repetitive transgenes occurs in *C. elegans* (Kelly et al. (1997)) and that silencing of the original Mos1 transposon extra-chromosomal array would eventually occur and decrease mutation rate (Williams et al. (2005); Vallin et al. (2012)).

Assuming that somatic sickness due to transposition can be resolved, the above leads to several proposals for increasing mutation rate. First (in order of difficulty), keeping P0 parents at a more Mos1-optimal temperature 25-28 °C instead of lowering the temperature to 20C °C after heat-shock. Second, using a different approach than microinjection, such as CRISPR or bombardment, to create a mutator strain with low copy numbers of transposon may counter-intuitively increase the mutation rate by preventing germline silencing. Finally, re-engineering the transposon injection plasmid, pBM020, to use the optimal Mos1 UTR at either end could also increase the transposition rate.

Another alternative would be to use an entirely different transposon altogether. The transposon *piggyBac* inserts into a random TTAA, has been shown to mobilize at 25 °C, be amenable to internal modification (Thibault et al. (2004)), has been optimized for high rates of transposition (Cadiñanos and Bradley (2007)), and has been shown to outperform Mos1 and other transposons in an *in vivo* model (Wu et al. (2006)). A *piggyBac* mutagenesis system has not be previously transformed into *C. elegans* but feasibility is likely due to its application in yeast (Li et al. (2011a)), *Drosophila* (Thibault et al. (2004)), and mammalian cells (Wu et al. (2006); Kong et al. (2010)). In order for *piggyBac* to work in *C. elegans*, its transposase would need to be modified similarly to Mos1; it would require an artificial intron as well as worm-specific promoters and 3' UTRs (Bessereau et al. (2001)). In addition to having a possibly higher mutation rate, a *piggyBac* system would allow us to put selectable markers within the transposon. Such markers would allow us to select against animals without transposon insertions, which currently account for more than 50% of the animals in our current model. Furthermore, a TnSeq protocol exists for the *piggyBac* transposon and the only modification required would be the AscI digest for removing array transposons (Bronner et al. (2016)).

5.2 Extending the CeTnSeq model

The simulation model from Chapter 2 can be further extended. For example, the mutation behavior (rate and TA-dinucleotide targets) of the Mos1 transposon could be replaced with that of another transposon such as piggyBac. The design of the starvation survival experiment is broadly applicable to any single selection experiment (e.g. heat shock, osmotic stress) and the empirical effect size distribution can be replaced. Finally, the Poisson error model used for sequencing error can be replaced with an empirical model appropriate for the library considered (e.g. PCR-error model). This ability to extend the power analysis model allows us to optimally design Ce-TnSeq experiment *in silico* before experimentation.

5.3 MIPs for RADseq

Like CeTnSeq, we required our RADseq reads to be representative of the underlying molecule counts in the original sample. We found evidence of noise in our RADseq reads which we attribute to PCR amplification noise. To address this, future work will use molecular inversion probes (MIPs) instead of RADseq for population sequencing. MIPs are single oligonucleotides whose 5' and 3' ends hybridize to a target sequence, and then a polymerase extension and ligation reaction cause the MIP to circularize and have a synthesized insert of the capture sequence (Hardenbol et al. (2003)). The background DNA is degraded with exonuclease, and the inserts of the circularized MIPs are PCR amplified and sequenced. MIPs are also designed with unique random barcodes to allow the discerning of PCR duplicates. This approach should give us the advantages of targeted sequencing (such as in RADseq) with additional fidelity in the representativeness of the sequencing library.

Bibliography

Ahringer, J. (2006), "Reverse genetics," *WormBook : the online review of C. elegans biology*, pp. 1–43.

Andersen, E., Gerke, J., and Shapiro, J. (2012), "Chromosome-scale selective sweeps shape Caenorhabditis elegans genomic diversity," *Nature . . .*, 44, 285–290.

Andersen, E. C., Shimko, T. C., Crissman, J. R., Ghosh, R., Bloom, J. S., Seidel, H. S., Gerke, J. P., and Kruglyak, L. (2015), "A Powerful New Quantitative Genetics Platform, Combining Caenorhabditis elegans High-Throughput Fitness Assays with a Large Collection of Recombinant Strains," *G3*, 5, 911–920.

Augé-Gouillou, C., Hamelin, M.-H., Demattei, M.-V., Periquet, M., and Bigot, Y. (2001), "The wild-type conformation of the Mos-1 Inverted Terminal Repeats is suboptimal for transposition in bacteria," *Molecular Genetics and Genomics*, 265, 51–57.

Baird, N. a., Etter, P. D., Atwood, T. S., Currey, M. C., Shiver, A. L., Lewis, Z. a., Selker, E. U., Cresko, W. a., and Johnson, E. a. (2008), "Rapid SNP discovery and genetic mapping using sequenced RAD markers." *PloS one*, 3, e3376.

Barquist, L., Mayho, M., Cummins, C., Cain, A. K., Boi, C., Page, A. J., Langridge, G. C., Quail, M. A., Jacqueline, A., and Parkhill, J. (2016), "The TraDIS toolkit : sequencing and analysis for dense transposon mutant libraries," *Bioinformatics*.

Baugh, L. R., Demodena, J., and Sternberg, P. W. (2009), "RNA Pol II accumulates at promoters of growth genes during developmental arrest." *Science (New York, N.Y.)*, 324, 92–4.

Bessereau, J. L., Wright, a., Williams, D. C., Schuske, K., Davis, M. W., and Jorgensen, E. M. (2001), "Mobilization of a Drosophila transposon in the Caenorhabditis elegans germ line." *Nature*, 413, 70–4.

Boulin, T. and Bessereau, J.-L. (2007), "Mos1-mediated insertional mutagenesis in Caenorhabditis elegans." *Nature protocols*, 2, 1276–87.

Bronner, I. F., Otto, T. D., Zhang, M., Udenze, K., Wang, C., Quail, M. A., Jian, R. H. Y., Adams, J. H., and Rayner, J. C. (2016), "Quantitative Insertion-site

Sequencing (QIseq) for high throughput phenotyping of transposon mutants," *Genome Research*.

Bryan, G., Garza, D., and Hartl, D. (1990), "Insertion and Excision of the Transposable Element Mariner in Drosophila," *Genetics*.

Cadiñanos, J. and Bradley, A. (2007), "Generation of an inducible and optimized piggyBac transposon systemy," *Nucleic Acids Research*, 35.

Cook, D. E., Zdraljevic, S., Roberts, J. P., and Andersen, E. C. (2016), "CeNDR , the Caenorhabditis elegans natural diversity resource," *Nucleic Acids Research*, pp. 1–8.

Dejesus, M. A., Ambadipudi, C., Baker, R., Sassetti, C., and Ioerger, T. R. (2015), "TRANSIT - A Software Tool for Himar1 TnSeq Analysis," *PLoS Computational Biology*, pp. 1–17.

Dembek, M., Barquist, L., Boinett, C. J., Cain, A. K., Mayho, M., Lawley, T. D., Fairweather, N. F., and Fagan, R. P. (2015), "High-Throughput Analysis of Gene Essentiality and Sporulation in Clostridium difficile," *mBio*, 6, 1–13.

Desfeux, C., Clough, S. J., and Bent, A. F. (2000), "Female reproductive tissues are the primary target of Agrobacterium-mediated transformation by the Arabidopsis floral-dip method." *Plant physiology*, 123, 895–904.

Frøkjær-Jensen, C., Davis, M. W., Ailion, M., and Jorgensen, E. M. (2012), "Improved Mos1-mediated transgenesis in C. elegans." *Nature methods*, 9, 117–8.

Gaertner, B. E. and Phillips, P. C. (2010), "Caenorhabditis elegans as a platform for molecular quantitative genetics and the systems biology of natural variation," *Genetics Research*, pp. 331–348.

Gallagher, L. A., Shendure, J., and Manoil, C. (2011), "Genome-Scale Identification of Resistance Functions in Pseudomonas aeruginosa Using Tn-seq," *mBio*, 2, 1–8.

Gawronski, J. D., Wong, S. M. S., Giannoukos, G., Ward, D. V., and Akerley, B. J. (2009), "Tracking insertion mutants within libraries by deep sequencing and a genome-wide screen for Haemophilus genes required in the lung," *Proc. Natl. Acad. Sci. U.S.A.*, 106, 16422–16427.

Goodman, A. L., Wu, M., and Gordon, J. I. (2012), "Identifying microbial fitness determinants by Insertion Sequencing (INSeq) using genome-wide transposon mutant libraries," *Nature Protocols*, 6, 1969–1980.

Hardenbol, P., Banér, J., Jain, M., Nilsson, M., Namsaraev, E. a., Karlin-Neumann, G. a., Fakhrai-Rad, H., Ronaghi, M., Willis, T. D., Landegren, U., and Davis, R. W. (2003), "Multiplexed genotyping with sequence-tagged molecular inversion probes." *Nature biotechnology*, 21, 673–8.

Hashimshony, T., Wagner, F., Sher, N., and Yanai, I. (2012), "CEL-Seq: Single-Cell RNA-seq by Multiplexed Linear Amplification," *Cell Reports*, 2, 666–673.

Jorgensen, E. M. and Mango, S. E. (2002), "The art and design of genetic screens: caenorhabditis elegans." *Nature reviews. Genetics*, 3, 356–69.

Kelly, W. G., Xu, S., Montgomery, M. K., and Fire, A. (1997), "Distinct requirements for somatic and germline expression of a generally expressed Caenorhabditis elegans gene," *Genetics*, 146, 227–238.

Kenyon, C., Chang, J., Gensch, E., Runder, A., and Tabtlang, R. (1993), "A C. elegans mutant that lives twice as long as wild type," *Nature*, 366, 461–464.

Kivioja, T., Vähärautio, A., Karlsson, K., Bonke, M., Enge, M., Linnarsson, S., and Taipale, J. (2012), "Counting absolute numbers of molecules using unique molecular identifiers," *Nature Methods*, 9, 3–7.

Kong, J., Wang, F., Brenton, J. D., and Adams, D. J. (2010), "Slingshot: A PiggyBac based transposon system for tamoxifen-inducible 'self-inactivating' insertional mutagenesis," *Nucleic Acids Research*, 38, 1–9.

Langridge, G. C., Phan, M.-d., Turner, D. J., Perkins, T. T., Parts, L., Haase, J., Charles, I., Maskell, D. J., Peters, S. E., Dougan, G., Wain, J., Parkhill, J., and Turner, A. K. (2009), "Simultaneous assay of every Salmonella Typhi gene using one million transposon mutants," *Genome Research*, pp. 2308–2316.

Li, H. and Durbin, R. (2009), "Fast and accurate short read alignment with Burrows Wheeler transform," *Bioinformatics*, 25, 1754–1760.

Li, J., Zhang, J. M., Li, X., Suo, F., Zhang, M. J., Hou, W., Han, J., and Du, L. L. (2011a), "A piggyBac transposon-based mutagenesis system for the fission yeast Schizosaccharomyces pombe," *Nucleic Acids Research*, 39.

Li, Z., Vizeacoumar, F. J., Bahr, S., Li, J., Warringer, J., Vizeacoumar, F. S., Min, R., Vandersluis, B., Bellay, J., Devit, M., Fleming, J. a., Stephens, A., Haase, J., Lin, Z.-Y., Baryshnikova, A., Lu, H., Yan, Z., Jin, K., Barker, S., Datti, A., Giaever, G., Nislow, C., Bulawa, C., Myers, C. L., Costanzo, M., Gingras, A.-C., Zhang, Z., Blomberg, A., Bloom, K., Andrews, B., and Boone, C. (2011b), "Systematic exploration of essential yeast gene function with temperature-sensitive mutants." *Nature biotechnology*, 29, 361–7.

Lohe, A. and Hartl, D. (2002), "Efficient mobilization of mariner in vivo requires multiple internal sequences," *Genetics*, 526, 519–526.

Martin, E., Alvarez, T., Bessou, C., Hauser, O., Sookhareea, S., Labouesse, M., and Segalat, L. (2002), "Note Identification of 1088 New Transposon Insertions of

Caenorhabditis elegans : A Pilot Study Toward Large-Scale Screens," *Genetics*, 162, 521–524.

Mello, C. and Fire, A. (1995), "DNA transformation," *Methods Cell Biol*, 48.

Merritt, C., Rasoloson, D., Ko, D., and Seydoux, G. (2008), "3' UTRs are the primary regulators of gene expression in the C. elegans germline," *Current Biology*, 18, 1476–1482.

Norrander, J., Kempe, T., and Messing, J. (1983), "Construction of improved M13 vectors using oligodeoxynucleotide-directed mutagenesis," *Gene*, 1, 101–106.

Page, D. R. and Grossniklaus, U. (2002), "The Art and Design of Genetic Screens: Arabidopsis Thaliana," *Nature Reviews Genetics*, 3, 124–136.

Pritchard, J. R., Chao, M. C., Davis, B. M., Baranowski, C., Zhang, Y. J., Rubin, E. J., and Waldor, M. K. (2014), "ARTIST : High-Resolution Genome-Wide Assessment of Fitness Using Transposon-Insertion Sequencing," *PLoS Genetics*, 10.

Ramsköld, D., Luo, S., Wang, Y.-c., Li, R., Deng, Q., Faridani, O. R., Daniels, G. A., Khrebtukova, I., Loring, J. F., Laurent, L. C., Schroth, G. P., and Sandberg, R. (2012), "Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells," *Nature Biotechnology*, 30.

Rinaldi, G., Eckert, S. E., Tsai, I. J., Suttiprapa, S., Kines, K. J., Tort, F., Mann, V. H., Turner, D. J., Berriman, M., and Brindley, P. J. (2012), "Germline Transgenesis and Insertional Mutagenesis in Schistosoma mansoni Mediated by Murine Leukemia Virus," *PLoS Pathogens*, 8, 1–13.

Sassetti, C. M., Boyd, D. H., and Rubin, E. J. (2001), "Comprehensive identification of conditionally essential genes in mycobacteria," *Proc. Natl. Acad. Sci. U.S.A.*, 98.

Schmitt, M. W., Kennedy, S. R., Salk, J. J., Fox, E. J., Hiatt, J. B., and Loeb, L. a. (2012), "Detection of ultra-rare mutations by next-generation sequencing." *Proceedings of the National Academy of Sciences of the United States of America*, 109, 14508–13.

Shaye, D. D. and Greenwald, I. (2011), "OrthoList : A Compendium of C . elegans Genes with Human Orthologs," *PLoS ONE*, 6.

Shuman, H. and Silvahy, T. (2003), "The art and design of genetic screens: Escherichia coli," *Nature reviews. Genetics*, 3, 176–88.

Sinzelle, L., Jégot, G., Brillet, B., Rouleux-Bonnin, F., Bigot, Y., and Augé-Gouillou, C. (2008), "Factors acting on Mos1 transposition efficiency." *BMC molecular biology*, 9, 106.

Solaimanpour, S., Sarmiento, F., and Mrázek, J. (2015), "Tn-Seq Explorer : A Tool for Analysis of High- Throughput Sequencing Data of Transposon Mutant Libraries," *PLoS ONE*, pp. 1–15.

St Johnston, D. (2002), "The Art and Design of Genetic Screens: Drosophila Melanogaster," *Nature Reviews Genetics*, 3, 176–188.

Stringham, E. G., Dixon, D. K., Jones, D., and Candido, E. P. (1992), "Temporal and spatial expression patterns of the small heat shock (hsp16) genes in transgenic Caenorhabditis elegans," *Mol Biol Cell*, 3, 221–233.

Sulston, J. E., Schierenberg, E., White, J. G., and Thomson, J. N. (1983), "The embryonic cell lineage of the nematode Caenorhabditis elegans," *Developmental Biology*, 100, 64–119.

Tas, H., Nguyen, C. T., Patel, R., Kim, N. H., and Kuhlman, T. E. (2015), "An integrated system for precise genome modification in Escherichia coli," *PLoS ONE*, 10, 1–19.

Thibault, S. T., Singer, M. A., Miyazaki, W. Y., Milash, B., Dompe, N. A., Singh, C. M., Buchholz, R., Demsky, M., Fawcett, R., Francis-Lang, H. L., Ryner, L., Cheung, L. M., Chong, A., Erickson, C., Fisher, W. W., Greer, K., Hartouni, S. R., Howie, E., Jakkula, L., Joo, D., Killpack, K., Laufer, A., Mazzotta, J., Smith, R. D., Stevens, L. M., Stuber, C., Tan, L. R., Ventura, R., Woo, A., Zakrajsek, I., Zhao, L., Chen, F., Swimmer, C., Kopczynski, C., Duyk, G., Winberg, M. L., and Margolis, J. (2004), "A complementary transposon tool kit for Drosophila melanogaster using P and piggyBac," *Nat Genet*, 36, 283–287.

Tosi, L. R. and Beverley, S. M. (2000), "cis and trans factors affecting Mos1 mariner evolution and transposition in vitro, and its potential for functional genomics." *Nucleic acids research*, 28, 784–90.

Vallin, E., Gallagher, J., Granger, L., Martin, E., Belougne, J., Maurizio, J., Duverger, Y., Scaglione, S., Borrel, C., Cortier, E., Abouzid, K., Carre-Pierrat, M., Gieseler, K., Ségalat, L., Kuwabara, P. E., and Ewbank, J. J. (2012), "A genome-wide collection of Mos1 transposon insertion mutants for the C. elegans research community." *PloS one*, 7, e30482.

van Luenen, H. G., Colloms, S. D., and Plasterk, R. H. (1993), "Mobilization of quiet, endogenous Tc3 transposons of Caenorhabditis elegans by forced expression of Tc3 transposase." *The EMBO Journal*, 12, 2513–2520.

van Opijnen, T. and Camilli, A. (2013), “Transposon insertion sequencing: a new tool for systems-level analysis of microorganisms.” *Nature reviews. Microbiology*, 11, 435–42.

van Opijnen, T., Bodi, K. L., and Camilli, A. (2009), “Tn-seq: high-throughput parallel sequencing for fitness and genetic interaction studies in microorganisms.” *Nature methods*, 6, 767–72.

Voutev, R. and Hubbard, E. J. A. (2008), “A ”FLP-out” system for controlled gene expression in Caenorhabditis elegans,” *Genetics*, 180, 103–119.

Wang, H., Liu, J., Gharib, S., Chai, C. M., Schwarz, E. M., Pokala, N., and Sternberg, P. W. (2016), “cGAL, a temperature-robust GAL4UAS system for Caenorhabditis elegans,” *Nature Methods*, 14.

Weigel, D., Ahn, J. H., Blázquez, M. A., Borevitz, J. O., Sioux, K., Fankhauser, C., Ferrándiz, C., Kardailsky, I., Elizabeth, J., Neff, M. M., Nguyen, J. T., Sato, S., Wang, Z.-y., Dixon, R. A., Harrison, M. J., Lamb, C. J., Yanofsky, M. F., Weigel, D., Ahn, J. H., Blazquez, M. A., Borevitz, J., Christensen, S. K., Fankhauser, C., Ferrandiz, C., Kardailsky, I., Malancharuvil, E. J., Neff, M. M., Nguyen, J. T., Sato, S., Wang, Z.-y., Xia, Y., Dixon, R. A., Harrison, M. J., Lamb, C. J., Yanofsky, M. F., and Chory, J. (2000), “Activation Tagging in Arabidopsis,” *Plant Physiology*, 122, 1003–1013.

Williams, D. C., Boulin, T., Ruaud, A.-F., Jorgensen, E. M., and Bessereau, J.-L. (2005), “Characterization of Mos1-mediated mutagenesis in Caenorhabditis elegans: a method for the rapid identification of mutated genes.” *Genetics*, 169, 1779–85.

Wu, S. C.-Y., Meir, Y.-J. J., Coates, C. J., Handler, A. M., Pelczar, P., Moisyadi, S., and Kaminski, J. M. (2006), “piggyBac is a flexible and highly active transposon as compared to sleeping beauty, Tol2, and Mos1 in mammalian cells.” *Proceedings of the National Academy of Sciences of the United States of America*, 103, 15008–13.

Yu, D., Huber, W., and Vitek, O. (2013), “Gene expression Shrinkage estimation of dispersion in Negative Binomial models for RNA-seq experiments with small sample size,” *Bioinformatics*, 29, 1275–1282.

Biography

Bradley Thomas Moore was born May 29, 1983, in Milwaukee, WI. He received his high school diploma in 2001 from Archbishop Moeller High School. He received his Bachelor of Science in Computer Science (*c*um laude) from The Ohio State University in 2003. He received his Master of Science degree in Computer Science from The Ohio State University in 2004 under Paul Sivilotti (M.S. thesis: Plausible Clocks with Bounded Inaccuracy). He received his Doctor of Philosophy in Computational Biology and Bioinformatics from Duke University in 2017 under Ryan Baugh.