NEXT-GENERATION SEQUENCING METHODS

FOR COMPLEX COMMUNITIES

by

ARIEL ELISE ROYALL

A DISSERTATION

Presented to the Department of Biology
and the Graduate School of the University of Oregon
in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy

June 2017

DISSERTATION APPROVAL PAGE

Student: Ariel Elise Royall

Title: Next-generation Sequencing Methods for Complex Communities

This dissertation has been accepted and approved in partial fulfillment of the requirements for the Doctor of Philosophy degree in the Department of Biology by:

Tory Herman	Chairperson
Eric Johnson	Advisor
John Postlethwait	Core Member
William Cresko	Core Member
Victoria DeRose	Institutional Representative

and

Scott L. Pratt	Dean of the Graduate School

Original approval signatures are on file with the University of Oregon Graduate School.

Degree awarded June 2017

DISSERTATION ABSTRACT

Ariel Elise Royall

Doctor of Philosophy

Department of Biology

June 2017

Title: Next-generation Sequencing Methods for Complex Communities

Advances in sequencing technology have opened up the possibility of investigating complex communities, but deviations from homogeneity in a sample create challenges in generating and analyzing sequence data. There are two kinds of heterogeneous populations that are addressed in this dissertation: low-frequency sequence variants in a group of largely homogeneous cells and rare members in complex biological communities. It is important to be able to fully characterize the heterogeneity of a sample, as rare genetic variants may provide fuel for selection and rare members of a complex community can play critical roles. Thus, heterogeneity can have important biological roles in everything from ecological community structure to human disease development and progression.

In order to assess low-frequency mutations, Paired-End Low Error Sequencing (PELE-Seq) was used. With this method, mutations occurring at frequencies as low as 1 in 10,000 were identified, including some with transcriptional consequences.
To investigate rare members of a larger community, an enrichment method was developed to sequence transcripts from host-associated bacteria. Rather than having to

sequence the abundant zebrafish host RNA, the enrichment protocol allowed even very minor members of the community to be efficiently sequenced, enabling a first look at the gene expression changes during colonization.

This dissertation includes previously published and unpublished material.

CURRICULUM VITAE

NAME OF AUTHOR: Ariel Elise Royall

GRADUATE AND UNDERGRADUATE SCHOOLS ATTENDED:

University of Oregon, Eugene
University of Texas, Austin

DEGREES AWARDED:

Doctor of Philosophy, Biology, 2017, University of Oregon
Bachelor of Science, Biochemistry, 2011, University of Texas

AREAS OF SPECIAL INTEREST:

Genomics
Bioinformatics

PROFESSIONAL EXPERIENCE:

Undergraduate Researcher, Marcotte Lab, University of Texas at Austin, 2009-2011

Undergraduate Researcher, Herzog Lab, University of Texas Medical Branch at Galveston, 2009

GRANTS, AWARDS, AND HONORS:

META Center for Systems Biology grant (NIH), University of Oregon, 2015-2017

PUBLICATIONS:

Efficient transcriptome profiling of host-associated bacteria. **Ariel Royall**, Catherine Pohl Robinson, Karen Guillemin, Eric Johnson. In Progress.

High-Specificity Detection of Rare Alleles with Paired-End Low Error Sequencing (PELE-Seq) Jessica L. Preston; **Ariel E. Royall**; Melissa A Randel; Kristin L Sikkink; Patrick C Phillips; Eric A Johnson. BMC Genomics. June 2016.

A proteomic survey of widespread protein aggregation in yeast. O'Connell JD, Tsechansky M, **Royall A**, Boutz DR, Ellington AD, Marcotte EM. Mol Biosyst. 2014 Apr;10(4):851-61. doi: 10.1039/c3mb70508k. Epub 2014 Feb 3.

Monitoring bacterial resistance to chloramphenicol and other antibiotics by liquid chromatography electrospray ionization tandem mass spectrometry using selected reaction monitoring. Haag AM, Medina AM, **Royall AE**, Herzog NK, Niesel DW. J Mass Spectrom. 2013 Jun;48(6):732-9. doi: 10.1002/jms.3220.

ACKNOWLEDGMENTS

I want to express sincere appreciation to Eric Johnson for his mentorship and assistance throughout my graduate career. Paul Etter's guidance in all technical aspects of library prep were invaluable. I also want to thank current and past members of the Johnson lab, Jessica Preston, Nick Kamps-Hughes, Melissa Randel, and Jim Stapleton, for their support and advice. I am grateful for the investments of my committee: Tory Herman, John Postlethwait, Bill Cresko, Victoria DeRose, and Andy Berglund. Finally, I would like to thank Doug Turnbull and Maggie Weitzman from the Genomics and Cell Characterization Core Facility for their assistance with all my projects. The investigation was supported in part by the META Center for Systems Biology grant (NIH).

TABLE OF CONTENTS

Chapter Page

LIST OF FIGURES

Figure Page

LIST OF TABLES

CHAPTER I

INTRODUCTION

Advances in sequencing technology have made possible once formidable tasks such as whole genome sequencing (Mardis, 2016). However, even small deviations from homogeneity create problems for the analysis of sequence data. For example, heterozygous alleles in a diploid organism can be mistaken for sequence error and vice versa (Wall, 2014). Advances in sequencing technology presented here have opened up possibilities for investigating complex communities that would previously have been obscured by errors and noise. There are two scenarios in which complex populations provide significant sequencing challenges that are addressed in this dissertation: sequence variants in a group of largely homogeneous cells and rare members in complex biological associations. It is important to be able to characterize these communities, as they play important biological roles in everything from ecological community structure to human disease development and progression.

Rare sequence variant within a population

A population of otherwise heterogeneous cells may have low frequency variation, as in tumors. Rare sequences within a population can be biologically relevant as they may provide an advantage to certain cells and can be selected for in such disease processes such as tumorigenesis and cancer progression. They are difficult to characterize at a very low level, because they are often present at or below typical sequencing error rates.

Two types of rare genetic variation were characterized. First, the presence of low-frequency variants was assessed in mitochondrial genomes in a tumor sample. Many copies of the mitochondrial genome are present within each cell, of which a few have a different genotype from the dominant genotype (Payne, 2013). Because of this low level but pervasive variation, there is a large pool of standing variation from which cancer cells may draw advantageous mutations. Cancer risk and outcomes have been correlated with mtDNA mutations (Chatterjee, 2006). Early assessment of these mutations could lead to more tailored treatment plans, but the sequencing error rate limits their detection. Here, a paired-end low error sequencing (PELE-Seq) is implemented to lower the error rate is used to identify functionally relevant mtDNA mutations down to 1 in 10,000. PELE-Seq is described in Chapter II, which has been published as Preston JL, Royall A, Randel MA, Sikkink KL, Phillips PC, and Johnson EA, 2015 "High-Specificity Next-Generation Sequencing of Minor Alleles with Paired-End Low Error Sequencing (PELE-Seq)"(*BMC Genomics*).

The second situation in which low-frequency variants were sequenced was assaying the generation of mutations in a zebrafish model of Fanconi Anemia. Fanconi Anemia (FA) is an autosomal recessive inherited DNA damage disorder resulting from the loss of one of the fanconi anemia proteins. These proteins are involved in two main complexes that mediate DNA damage repair. Previous studies have found significantly higher point mutation rated in cell line models of Fanconi Anemia (Araten, 2005). Experimental limitations previously prevented studying variants aside from large chromosomal rearrangements in patient samples or model organisms, leaving the

spectrum of somatic mutation in FA remains largely unexplored.By combining PELE-Seq with a longitudinal sampling scheme in fanconi anemia model zebrafish and multiple tissues from zebrafish and mice, new patterns of small scale mutation were uncovered.

Rare member of a larger, complex community

Rare members can play an important role in community health and maintenance as they can create products which affect the entire community. Vertebrate gut-associated bacteria have been shown to play essential roles in the health and development of their animal hosts, including facilitating digestion and nutrient acquisition, education and maturation of the immune system, and protection from pathogens (reviewed in Neish, 2009). Our understanding of these roles has been transformed by sequencing technologies that allow an unbiased look at the composition and activity of bacterial communities and the development of model animal systems for mechanistic studies into these intimate biological relationships. Traditional approaches to understanding host-associated bacterial communities such as 16S sequencing provide taxonomic data, but do not assess total gene content of the community or the genome-wide expression data. Transcriptomics provides information about both gene content and the relative activity of the genes within and across conditions. Previous work has shown the intestinal environment to be highly dynamic and that the spatial structuring of different bacterial species within the gut undergoes dynamic responses to the changing environment (Wiles, 2016). Tying transcriptional changes to specific phenotypes enriches our understanding of the genetic underpinnings of those phenotypes and gives new

insight on ways to manipulate host-associated microbes for specific goals. The large amount of contaminating host RNA and the short bacterial RNA half-life has made the transcriptome of host associated microbes inaccessible. This dissertation describes a method for enriching for the transcriptome of host-associated bacteria in Chapter IV.

CHAPTER II

PAIRED-END LOW ERROR SEQUENCING

This previously published co-authored material can be found here: Preston JL, Royall AE, Randel MA, Sikkink KL, Phillips PC, Johnson EA. High-specificity detection of rare alleles with Paired-End Low Error Sequencing (PELE-Seq). BMC Genomics. 2016;17(1):1543–21. The material was co-authored by Jessica Preston, Melissa Randel, Kristen Sikkink, Patrick Phillips, Eric Johnson and myself and published in 2016 in BMC Genomics. This work is included in my thesis as I contributed to the assay design and validation. I was also involved in designing and validating data analysis pipelines. Additionally, this work is critical to understanding subsequent chapters (III and IV).

BACKGROUND

Populations with high levels of genetic heterogeneity are able to evolve rapidly through natural selection, for example providing the basis for drug resistance in populations of microbes, viruses, and tumor cells (Kaiser 2013, Bahtia 2012, Modi 2013). In order to understand how these heterogeneous populations evolve in response to selection, it is important to be able to characterize the full catalog of genetic variation present in the population, including *de novo* mutations and minor alleles.

The reduced cost of DNA sequencing has powered the wide-scale discovery of functional and disease-causing single nucleotide polymorphisms (SNPs) and genomic

regions under selection (Hohenlohe 2010). However, the current high error rate (~1%) leads to the generation of millions of sequencing errors in a single experiment. Thus, when attempting to sequence *de novo* mutations or genetically heterogeneous populations, it is challenging to distinguish between sequencing errors and true rare genetic variants (Nielsen 2011, Marcais 2015, Schlossnig 2013, Kircher 2010).

Sequencing error reduction through the use of overlapping read pairs (ORPs) has been described previously by Chen-Harris *et al.,* who showed that the use of overlapping paired-end reads dramatically reduces the occurrence of sequencing errors (Goto 2011). PELE-Seq improves on the ORP method by incorporating dual-barcoding to filter out many types of PCR errors and library preparation artifacts.

The PELE-Seq method is simple to use, compatible with most sequencing libraries, and doesn't require the use of special reagents. The PELE-Seq error-reduction method is based on two principles. First, sequencing errors can be removed by sequencing each DNA molecule twice with overlapping reads and merging the reads into overlapping read pairs (ORPs). Any bases that are mismatched in the two sequences are excluded from the final SNP calling analysis. Second, PCR errors and library preparation artifacts are reduced through the use of a dual-barcoding system, which can be used to generate information about the number of independent occurrences of a genetic variant in a DNA sequencing library. The PELE-Seq variant calling analysis pipeline incorporates information from the barcoding data as well as the overlapping read pair data, and is

customized to allow for highly sensitive detection of rare polymorphisms without the losses in specificity compared to standard methods of DNA sequencing.

We applied the PELE-Seq method to sequence rare alleles in a wild population of *Caenorhabditis remanei* nematode worms. *C. remanei* are highly heterogeneous, non-hermaphroditic nematode worms that are amenable to studies investigating the genetic basis of the response to natural selection (Osvaldo 2010). In this study, we sampled the genome of an ancestral population originating from 26 wild mating pairs from Toronto, Ontario that were lab-propagated for a total of 34 generations. We show that PELE-Seq can detect changes in the rare allele frequencies between the genomes of the wild and lab-adapted populations, and that PELE-Seq can detect putative low-frequency *de novo* mutations that appear in the laboratory adapted population.

RESULTS

PELE-Seq Library Preparation and Data Analysis

PELE-Seq improves the specificity of standard SNP calling methods by reducing the occurrence of false-positive sequencing errors in the data. An overview of the PELE-Seq method is illustrated in Figure 2-1.

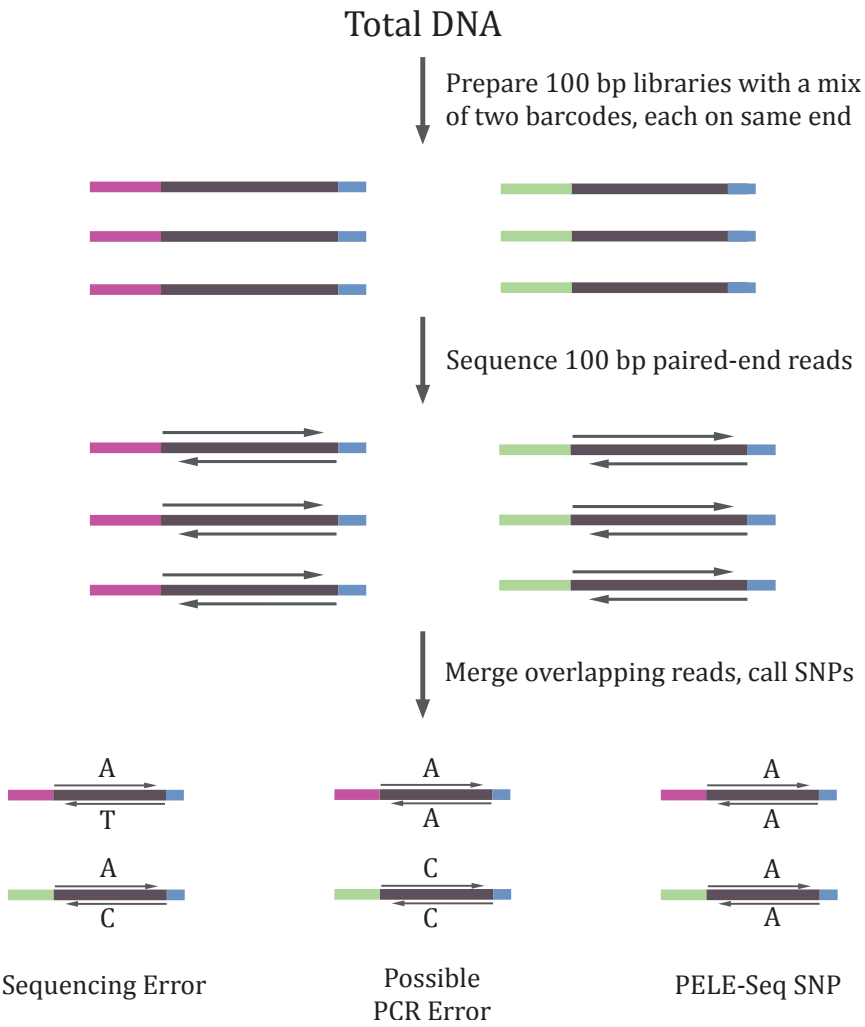

Figure 2-1. **The PELE-Seq method of rare variant calling**. DNA libraries with a 100bp insert size are paired-end sequenced using 100bp reads, generating an overlap region of approximately 100bp. The overlapping reads are merged into a consensus sequence and mismatching bases are discarded. A mixture of two separate barcodes is ligated to each sample. In order to pass PELE-Seq quality filtering, SNPs must be present in both paired-end reads and with both barcodes.

PELE-Seq library preparation and analysis involves two separate error filtering steps which are combined during analysis:

1. Illumina 100 bp paired-end sequencing of short 100 bp DNA inserts is used to generate two completely overlapping paired-end reads from each DNA molecule. The overlapping paired-end reads are then merged into one high-quality consensus sequence. After trimming off the overhanging bases and filtering for high quality scores, the resulting consensus sequence has a much lower incidence of false positive SNPs compared to the non-overlapped reads.

2. PCR errors and library preparation artifacts are reduced through the use of a dual-barcoding system, which requires the presence of two independent occurrences of a variant. During library preparation, a two independent barcodes are ligated to the DNA molecules to be sequenced. Then, during data analysis, SNPs that are present with only a single barcode are excluded from the analysis, as they are potential PCR errors or library preparation artifacts.

PELE-Seq data analysis uses a multi-step variant calling approach to incorporate information from both the barcoding and the overlapping steps, without a large drop in sensitivity. Rare alleles are evaluated with the program LoFreq, which calls somatic variants using a Bonferroni-corrected *P*-value threshold of 0.05 (Chen-Harris 2011). Rare nucleotides are included in the final variant calling only if they pass two separate quality control steps: 1. The nucleotide is present in both overlapping sequence reads from a

single DNA molecule and is called as a SNP when variants are called from the merged reads. 2. The nucleotide is called as a SNP in two separate instances of high-sensitivity variant calling, once for each barcode file. The final outcome of the PELE-Seq analysis is a set of very high quality SNPs that have passed numerous quality control tests and filters.

PELE-Seq Specificity and Sensitivity

We first sought to empirically determine the specificity and sensitivity of the PELE-Seq variant calling method. We sequenced control *E. coli* DNA mixtures containing 64 known SNPs present at defined frequencies ranging from 0.1%-0.3%. The *E. coli* control DNA mixtures were generated using DNA from *E. coli* K12 substrain W3110 titrated into a much larger amount of DNA from *E. coli* B substrain Rel606. The K12 W3110 substrain of *E. coli* contains a SNP every ~117 bp compared to *E. coli* B substrain Rel606 (Costello 3012, Jeong 2009). The genome space sequenced was reduced to 14 kilobases by using Restriction-site Associated DNA Sequencing (RAD-Seq) to sequence only the 200 nucleotides flanking an SbfI restriction enzyme cut site (Hayashi 2006). SbfI cuts the sequence CCTGCAGG, which occurs ~70 times in the *E. coli* genome. We identified the control SNPs by sequencing the pure *E. coli* K12 substrain W3110 and comparing it to pure *E. coli* B substrain Rel606.

The identity and allele frequency of the *E. coli* SNPs in the control libraries was verified by sequencing to 25,000X average read depth (Table 2-1). The total read depth listed is that of the processed bam file used for SNP calling; for PELE-Seq data the

number of raw reads used to generate the final bam file is roughly 2.3 times this amount because of the overlapping stage of analysis. The rare alleles detected in the control libraries had allele frequencies ranging from 0.141-0.464% (1/200-1/710).

	Read Depth		Allele Frequency	
Library	mean	sd	mean	sd
1	26908	7357	0.003037	0.0007274
2	24182	9506	0.002284	0.0005316
3	33547	8079	0.002233	0.0005342
4	21631	3166	0.002128	0.0006200

Table 2-1. **Allele frequencies for known rare SNPs in control *E. coli* DNA mixtures.** Synthetic control *E. coli* libraries, labelled 1-4, were sequenced to an average read depth of 25,000X. The rare alleles detected in the control libraries had average allele frequencies ranging from 0.21-0.30% or 1/330-1/470 of total reads.

We found that PELE-Seq had high sensitivity with no false positive SNP calls when detecting rare SNPs above 0.2% allele frequency and with read depths below 30,000X (Figures 2-2,2-3). When detecting rare alleles known to be present at 0.3% frequency, PELE-Seq was able to correctly identify 22 out of the 64 total SNPs present with no false positives, while standard DNA-Seq methods with high base-quality (>Q30) identified 17 true SNPs, and had a false positive rate of 30%.

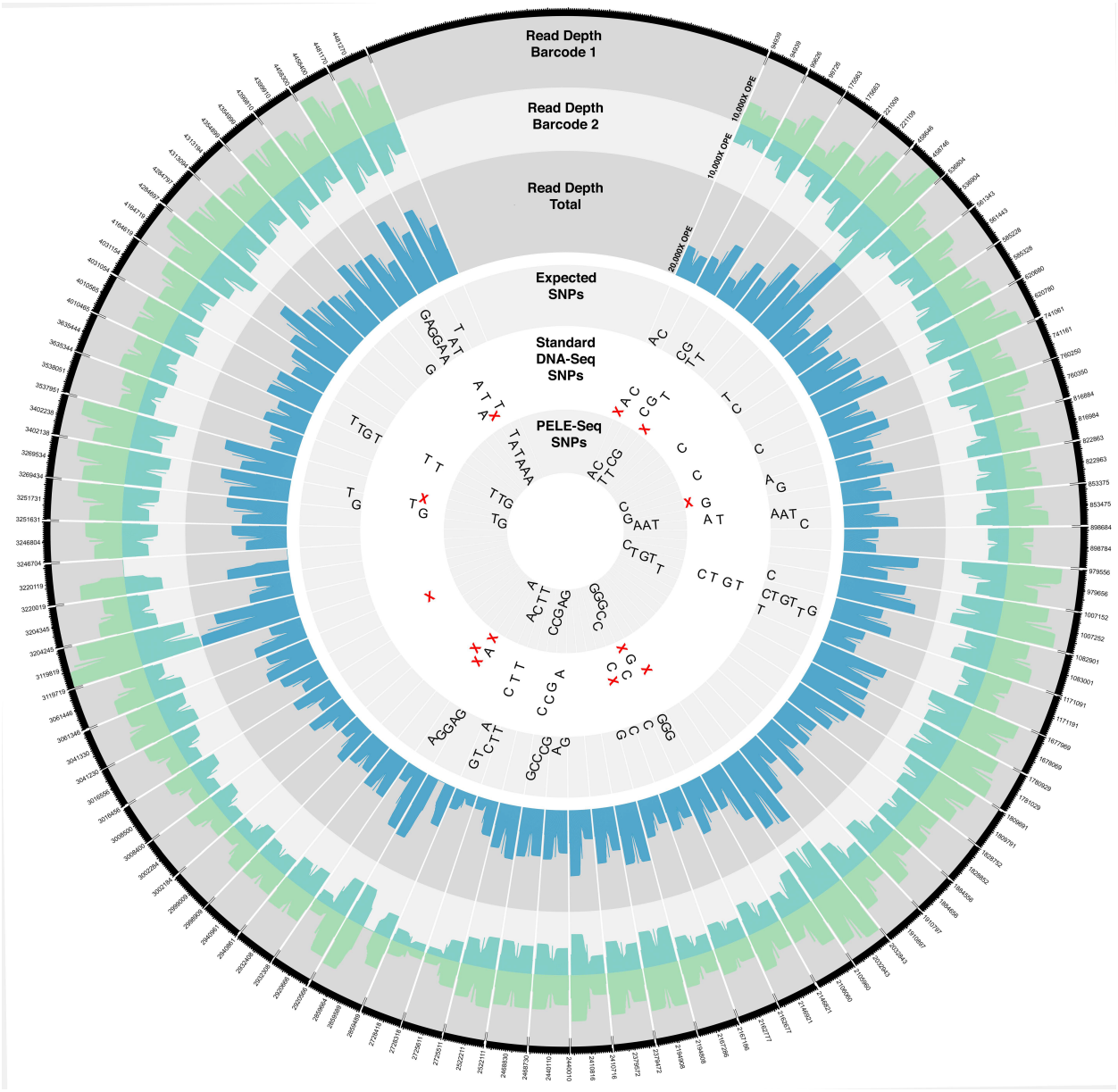

Figure 2-2. **Detecting SNPs present at 0.3% frequency in *E. coli* control libraries with PELE-Seq and standard DNA-Seq methods. .** Rare alleles present at 0.3% frequency in synthetic *E. coli* libraries were sequenced with PELE-Seq and standard DNA-Seq, at 20,000X average read depth. The read depths of the individual barcode files are plotted in blue, and the total read depth is plotted in green. The SNPs detected with PELE-Seq are plotted in the inner circle, and the Non-PELE SNPs are plotted in the next outer circle. False positive mutations are designated with a red "X". Of the 64 known SNPs present in the genome, PELE-Seq detected 22 mutations with 100% specificity, compared to 17 mutations and 70% specificity achieved with non-PELE methods.

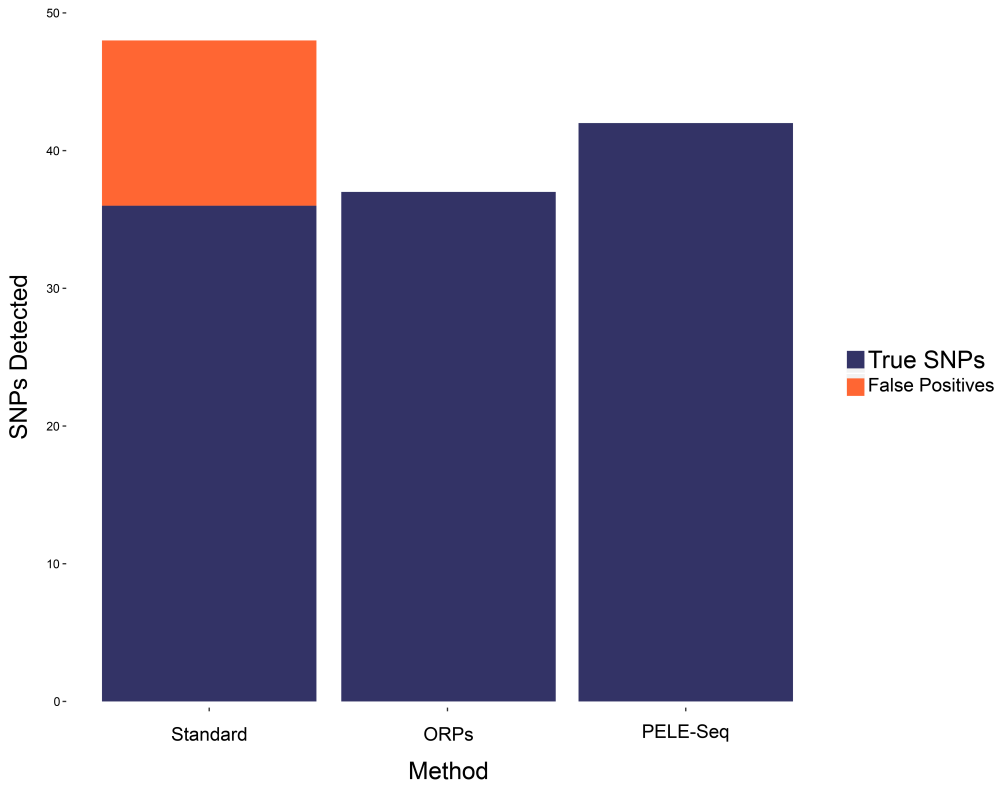

Figure 2-3. **PELE-Seq data has zero false positive SNPs and is more sensitive than standard DNA-Seq data.** Sequencing a control *E. coli* DNA library containing 64 rare SNPs present at 0.3% allele frequency with PELE-Seq at 20,0000X read depth produces 100% specific data, compared to 71% specificity achieved with traditional sequencing methods. Traditional Non-PELE sequencing of the control libraries resulted in 7 false positive mutations, compared to zero with the PELE-Seq method.

We compared the specificity of the PELE-Seq method to that of the previously developed "Overlapping Read Pair (ORP)" method of rare SNP detection in order to determine the benefit of using multiple barcodes and a custom analysis pipeline. When just overlapping read error correction was used, false positive SNP calls were made compared to the no false positives seen with PELE-Seq (Table 2-2).

Table 2 Rare SNPs identified using the PELE-Seq, ORP, and standard DNA-Seq methods, at various read depths

Average read depth per barcode	PELE positives	PELE false positives	ORP positives	ORP false positives	Standard positives	Standard false positives
1000	13	0	6	0	6	2
5000	19	0	18	0	24	7
10000	42	0	37	0	36	12
15000	36	0	32	0	35	13
18000	40	0	35	0	41	42

Table 2-2. Rare SNPs detected with PELE-Seq, standard DNA-Seq, and the ORP method. Rare alleles present at 0.3% frequency in synthetic *E. coli* libraries sequencing at 20,000X depth of coverage were sequenced with PELE-Seq, standard DNA-Seq, and the ORP method. PELE-Seq is more specific than standard DNA-Seq and the ORP method, with zero false positive SNPs detected.

Detection of rare and putative *de novo* mutations in wild and lab-adapted *C. remanei*

We applied PELE-Seq to track changes in the rare allele frequencies of a wild population of *C. remanei* nematode worms that was subjected to laboratory-adaptation. The ancestral (wild) *C. remanei* population originated from 26 mating pairs of nematodes that were expanded to a population of 1000+ individuals and then frozen within three generations [10]. A branch of this ancestral population was grown in the lab for 34 generations, during which time it was culled randomly to a population of 1000 individuals for each generation. The lab-adapted population was also subjected to 2 freezes and 9 bleach treatments (hatchoffs) during this time. The numerous selection events endured by the lab-reared nematodes are expected to lower genetic diversity of the population via drift and bottlenecking. Rare advantageous SNPs may also be selected for during the process of lab-adaptation.

To assess the changes in genetic diversity of the nematode population before and after lab-adaptation, DNA from the wild and laboratory-adapted populations of *C. remanei* worms was PELE-sequenced using PacI RAD-Seq. The PacI restriction enzyme cuts the sequence AATTAATT, which occurs 2044 times in the *C. remanei* caeRem3

genome. In order to further decrease the complexity of the genome, we performed an additional restriction enzyme digestion with NlaIII to destroy a portion of the RAD tags in the library. NlaIII cuts the sequence CATG, which is present on approximately 30% of the PacI RAD tags. The resulting genome space covered was approximately 300 kb, which was sequenced to an average of 2000X read depth.

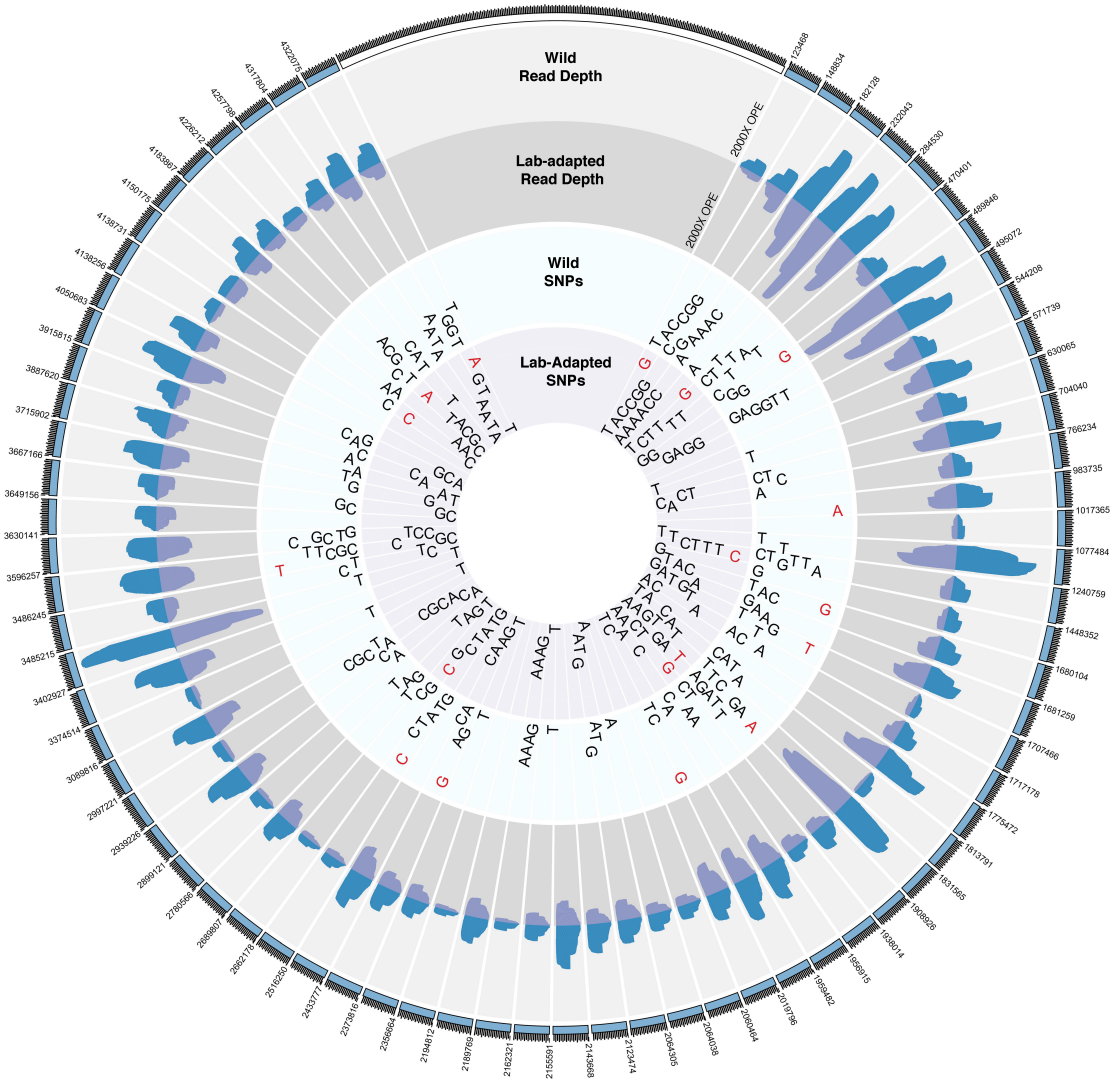

Figure 2-4. PELE-Sequencing of SNPs in wild and lab-adapted *C. remanei* populations. The inner yellow circle lists SNPs present in the lab-adapted population; the wild SNPs are listed in the blue circle. SNPs present in both the wild and lab-adapted populations are written with black letters. SNPs appearing in only the wild or lab-adapted populations are written with red letters.

We identified several differences between the SNPs present in the wild nematodes compared to those found in the lab-adapted population (Figure 2-4). We found SNPs present below 1% frequency that were unique to the wild or lab-adapted *C. remanei* populations, and the frequencies of some of these rare alleles changed dramatically during lab-adaptation.

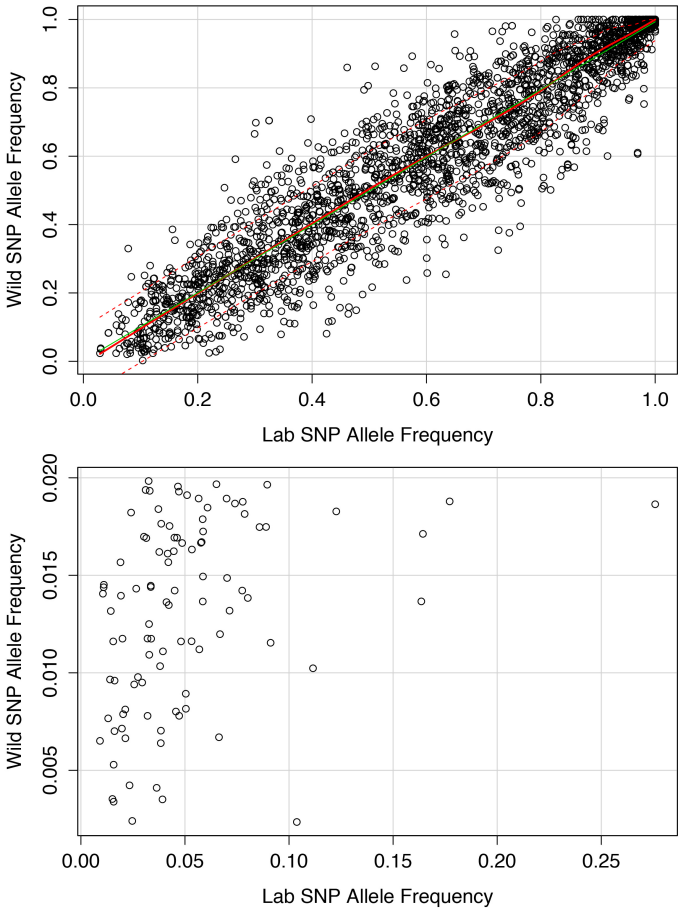

Figure 2-5. The allele frequencies of SNPs in the ancestral and lab-adapted populations of *C. remanei* worms. Each point represents a SNP in the genome. **Top)** Allele frequencies before and after lab-adaptation for all SNPs detected that are present in both populations. SNPs in the top left corner are less frequent in the lab-adapted worms; SNPs in the bottom right corner are more frequent in the lab-adapted worms. The estimated 0.25 and 0.75 quantiles of the square root of variance are shown for with the dashed red lines. **Bottom)** A zoom-in of allele frequencies for SNPs present below 1% in the wild *C. remanei* population, before and after lab-adaptation. Five minor alleles present below 1% in the wild population increased in frequency fivefold after lab adaptation. Only SNPs present in both populations are plotted.

By plotting the allele frequencies of each SNP before and after lab adaptation, it is possible to visualize the changes in the allele frequencies of minor alleles in a population undergoing a response to selection. The most dramatic changes in SNP allele frequencies were observed in the rare SNPs (Figure 2-5).

We identified 4658 PELE-quality SNPs present below 1% frequency in the ancestral *C. remanei* population, and 2541 PELE-quality SNPs present below 1% frequency in the lab-adapted population. Of the 4658 SNPs that were present below 1% the ancestral *C. remanei* population, 958 SNPs were still detected in the lab-adapted population, including 534 SNPs below 1% in the lab-adapted population. There were 14 SNPs that were found to increase in frequency at least tenfold in the lab-adapted population compared to the ancestral population (Table 2-3).

Position	Ref	Alt	AF Wild	Reads Wild	AF Lab	Reads Lab	Fold Change AF
4938079	A	C	0.0097	19	0.20	116	23
4938081	T	C	0.0086	17	0.19	115	20
31252148	G	A	0.0090	9	0.20	103	14
31487455	G	A	0.0095	31	0.18	257	17
33492880	G	A	0.0085	22	0.20	195	12
57798676	G	C	0.0098	21	0.13	144	19
76928211	G	C	0.0078	18	0.13	80	11
85765886	G	A	0.0092	34	0.11	311	14
103193682	A	G	0.0097	8	0.11	46	14
125627381	A	G	0.0083	34	0.11	268	14
125627408	A	G	0.0084	41	0.13	397	22
127488550	T	C	0.0082	37	0.12	252	23
127488619	G	A	0.0076	40	0.13	313	17
127723967	C	G	0.0023	31	0.10	747	16

Table 2-3. Rare SNPs in wild *C. remanei* that have increased after lab adaptation. Five SNPs present below 1% frequency in the wild *C. remanei* population increased in frequency at least 5x in the lab-adapted population.

A SNP was detected at position 127,723,967 of the caeRem3 (WUSTL) genome that had increased in frequency by 45X in the lab-adapted population. The number of reads containing this G>C transversion jumped from 31/13000 (0.2%) in the wild population to 750/7000 (10.5%). This SNP is located upstream of the promoter region of

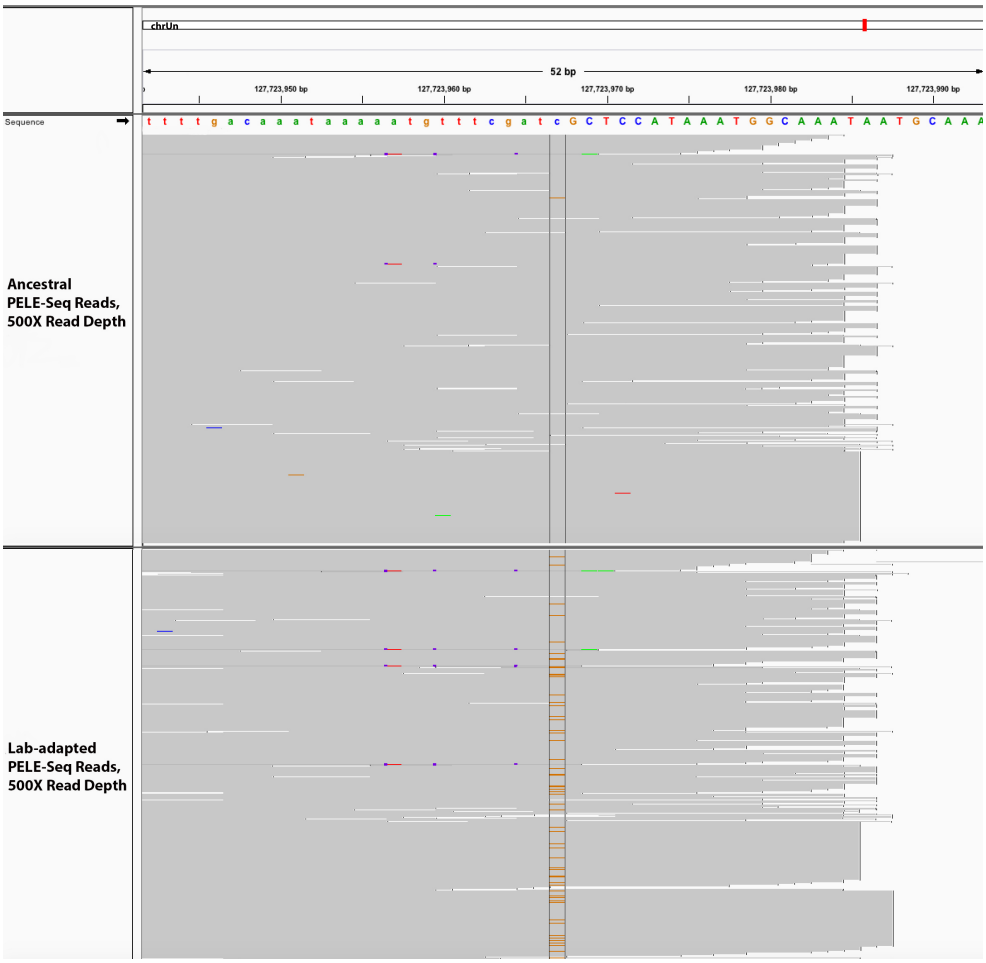

Figure 2-6. A RAD tag sequenced with PELE-Seq contains a SNP mapping to the promoter region of ugt-5. A rare SNP present at position 127,723,967 of the caeRem3 (WUSTL) genome maps to the predicted *C. elegans* gene ugt-5. The SNP increased in frequency by 44X after 34 generations of lab-adaptation. The UGT pathway is a major pathway responsible for the removal of drugs, toxins, and foreign substances. The top panel shows the reads from the ancestral (wild) population mapping to the caerem3 genome; the bottom panel shows the reads from the lab-adapted population. The non-reference SNP at position 127,723,967 is visible in orange.

a gene predicted by UCSC to be homologous to the *C. elegans* gene ugt-5, a UDP-Glucuronosyltransferase (Figure 2-6). The reads mapping to this SNP in IGV are shown.

The lab-adapted worms also contained rare SNP that were not detected in the wild population, including putative *de novo* mutations. We identified 287 rare variants that were present only in the lab-adapted *C. remanei* population. These rare alleles were called with extremely high stringency by removing any SNPs that were called with either barcode file in the wild population from the analysis. The rare alleles appearing only in the lab-adapted population are all present below 0.8% allele frequency and are distributed throughout the genome (Figure 2-7).

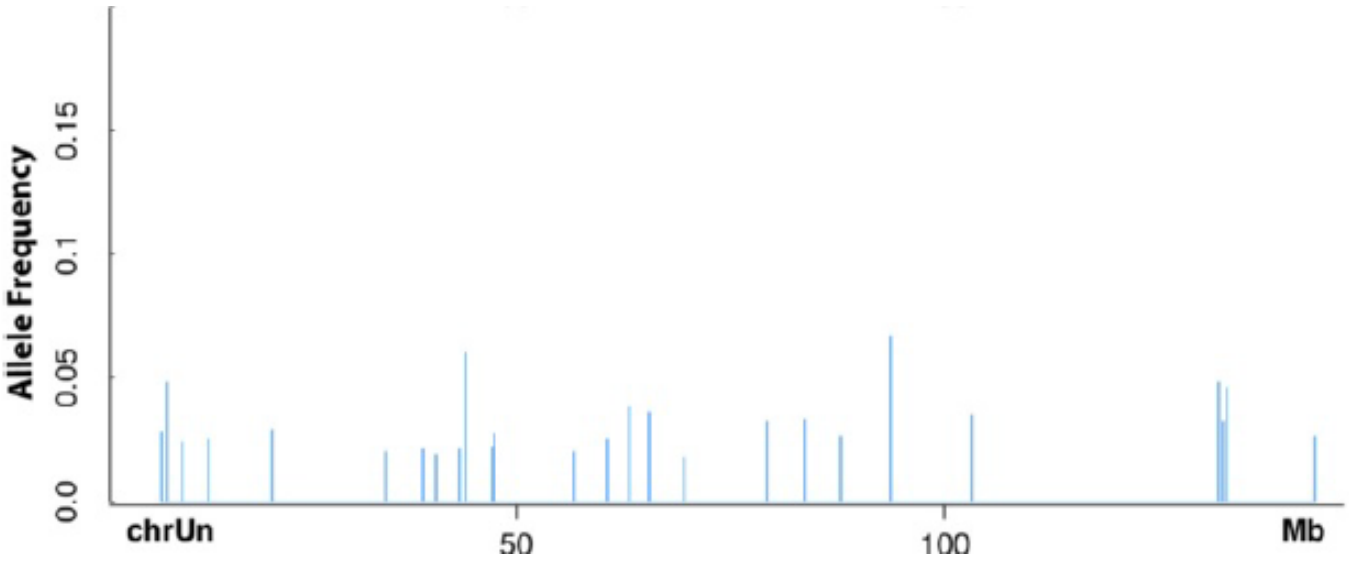

Figure 2-7. Allele frequencies and position of 49 mutations detected only in the lab-adapted *C. remanei* population with PELE-Seq. Each vertical line represents a single *SNP*; the height of the line is proportional to the allele frequency. The detected SNPs had allele frequencies ranging from 0.0021 to 0.0070. The UCSC caeRem3 genome from WUSTL is composed of a single artificial chromosome named chrUn that is 146 megabases (Mb) long.

DISCUSSION

Current genomic studies of genetically heterogeneous samples, such as *de novo* mutations in growing tumors or natural populations that are difficult to sequence as individuals, are hampered by the difficulty in distinguishing alleles at low frequency from the background of sequencing and PCR errors. We have developed a method of rare allele detection that mitigates both sequence and PCR errors called PELE-Seq. PELE-Seq was evaluated using synthetic *E. coli* populations and used to compare a wild *C. remanei* population to a lab-adapted population. Our results demonstrate the utility of the method and provide guidelines for optimal specificity and sensitivity when using PELE-Seq.

By using PELE-Seq, we increased the number of independent validations of a rare SNP by sequencing each molecule twice with overlapping paired-end reads and by calling each SNP twice through the use of multiple barcodes. The multiple PELE-Seq quality control steps result in genotype calls of low-frequency alleles with a false positive rate of zero, allowing for the specific detection of rare alleles in genetically heterogeneous populations.

We found that there is a window of sequencing depth that is ideal for detecting rare alleles when using PELE-Seq, and sequencing beyond this level will increase the probability of introducing false positive mutations due to PCR error. The ideal amount of coverage for a given library would depend on the specific PCR error rate of the method used to make the library. For our libraries, with an estimated PCR error rate of 0.05%, we found that the optimal level of read depth was around 25,000X coverage. Sequencing

below this level reduced the sensitivity of the method, while sequencing above this level lead to the appearance of PCR errors in the data that were present in both barcoded libraries.

Sequencing error reduction through the use of overlapping read pairs (ORPs) has been described previously by Chen-Harris *et al.,* who show that the use of overlapping paired-end data dramatically reduces the occurrence of sequencing errors in NGS data (Goto 2011). Their group concluded that PCR error is the dominant source of error for sequencing data with an Illumina quality score above Q30, which they estimate to be around 0.05%. PELE-Seq adds to the overlapping read pair method by incorporating dual barcodes to filter out the PCR errors. We have shown that the PELE-Seq method has fewer false positives than sequencing data generated with the ORP method alone in our libraries.

We have used PELE-Seq to identify rare alleles in a wild *C. remanei* population whose frequencies have increased dramatically as result of laboratory cultivation, and we identify putative *de novo* mutations that have arisen during laboratory adaptation of a wild nematode worm population. We identified a rare G > C transversion upstream of the promoter of ugt-5 that was increased in frequency 45X in the lab-adapted strain compared to the wild strain. UGT enzymes catalyze the addition of a glucuronic acid moiety onto xenobiotics and drugs to enhance their elimination. The UGT pathway is a major pathway responsible for the removal of most drugs, toxins, and foreign substances (Sikkink 2014). The striking increase in the frequency of this rare mutation after lab

adaptation suggests that the surrounding genomic region is under positive selection. One possibility is that a change in ugt-5 expression may confer a growth advantage on the laboratory-grown nematodes by increasing their ability to process and eliminate the bleach ingested during the hatchoff procedures. With PELE-Seq, it is possible to know that the ugt-5 SNP was present at a very low frequency in the wild population, and is not a *de novo* mutation. The SNPs detected only in the lab-adapted population were present at low frequencies, suggesting that pre-existing low-frequency minor alleles are the most useful source of genetic material available for *C. remanei* to respond to changes in the environment, as these alleles are readily available and don't need to be spontaneously generated. In general, this approach should be useful for detecting changes in rare allelic variants in so-called "evolve and reseq" experiments. In this study, we sampled only a very small fraction (~1/500) of the *C. remanei* genome with RAD-Seq, and discovered multiple instances of apparent selection taking place.

CONCLUSIONS

We have demonstrated that the PELE-Seq method of variant calling is highly specific at detecting rare SNPs found at below 1% of a population. There were zero instances of false positive SNPs called from control sequenced *E. coli* library containing known rare alleles present at known frequencies. Previously, the high error rate of NGS resulted in thousands of false-positive SNPs that were indistinguishable from true minor alleles. The PELE-Seq method makes it possible to know with certainty the identity of rare alleles in a genetically heterogeneous population, and to detect ultra-rare and putative *de novo* mutations that aren't present in an ancestral population. As a proof of principle, we have used PELE-Seq to identify rare mutations found in lab-adapted strains of *C. remane*i nematode worms. We identified a SNP in the lab-adapted worms that was increased in frequency more than 40X after 23 generations in the lab. This research demonstrates that model organisms grown in a laboratory can become genetically distinct from wild populations in a short period of time, and care must be taken when generalizing from conclusions drawn from research involving lab-reared organisms.

BRIDGE

Once an experimental and computational pipeline for distinguishing rare variant from sequencing error and noise was developed, we were interested in using this method to investigate two highly heterogeneous communities: tumor mitochondria and somatic cells in Fanconi Anemia.

CHAPTER III

VARIANTS IN TUMOR MITOCHONDRIAL DNA

A. Introduction

Mitochondrial DNA is present in 1,000 to 10,000 copies per cell, depending on cell type. The majority of these mitochondria will have identical, maternally inherited mitochondrial genomes. A small fraction of the mitochondrial of that cell will have an alternate genotype, which is called heteroplasty (Figure 3-1). In 2013, Payne *et al* reported mtDNA heteroplasmy in all samples analyzed, but all were present below 0.2%.

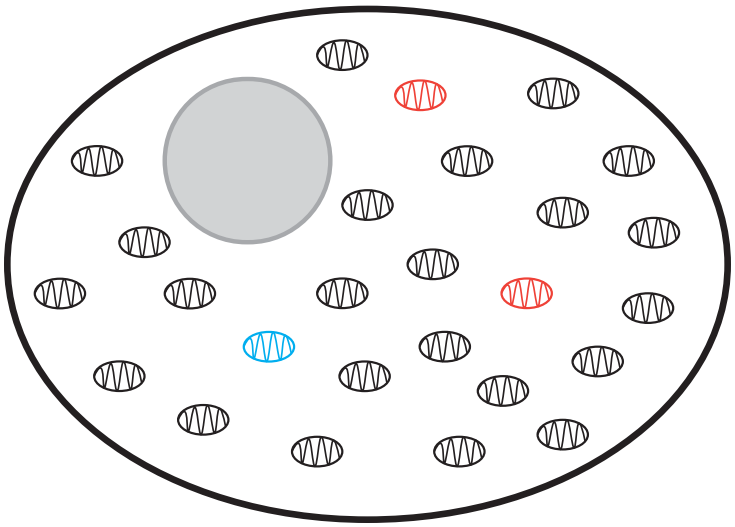

Figure 3-1. Heteroplasmy in Mitochondrial DNA. Each cell contains many copies of the mitochondrial genome, most of which have a single genotyle (black). However, a few mitochondria (usually less than 0.2%), have an alternate genotype (pink and blue).

Because of this low level but pervasive variation, there is a large pool of standing variation from which cancer cells may draw advantageous mutations. mtDNA mutations have been associated with many types of cancer and correlated with cancer outcomes (Chatterjee 2006). The mutations characterized thus far have typically been homoplasmic in nature and specific to cancerous tissue, suggesting that they arise early in

tumorigenesis or are selected for in early tumorigenesis from standing variation within the precancerous cells. Identifying these mutations early in diagnosis and treatment could lead to a better prognosis, however finding low level mutation is difficult with standard sequencing techniques. Investigating the low level variation in mitochondrial DNA becomes feasible with PELE-Seq. Here, PELE-Seq is employed to examine how mutations arise and propagate in tumor mitochondrial DNA.

B. Materials and Methods

Many techniques exist to examine polymorphisms between individuals, and within populations of individuals, but it remains difficult to describe low-level variation that arises during development, growth, tumorigenesis, etc. These ultra-rare alleles cannot be detected reliably by standard sequencing techniques, as their incidence is often below the sequencing error rate and thus they are indistinguishable from noise. Paired-End low error sequencing (PELE-Seq) provides a tool to detect alleles present down to 0.01% (Figure 2-1).

By incorporating short, 100bp inserts and multiple barcodes, PELE-Seq is able to detect alleles in this low range. DNA is made into sequencing libraries with an insert size equal to the read length. When these fragments are sequenced paired-end, the forward and reverse read cover the same stretch of DNA and each base is read twice. The probability of the same error occurring in both the forward and the reverse read becomes very low (theoretically 0.000001% for positions with phred-scaled qualities of 40 on both forward and reverse reads). PCR amplification steps in library preparation can create and amplify errors that later sequence as high quality alleles. To correct for these, each sample is given multiple barcodes. An allele must be present in both barcodes to be considered a true alternate allele. This way the probability of PCR errors being falsely called SNPs with PELE-Seq analysis becomes very low also (theoretically o.ooo4% with HF Phusion polymerase and 200bp fragments). This approach allows investigation of ultra-rare alleles in a variety of systems.

Sequencing reads do not typically fully overlap and individual samples generally have only one barcode, so a new data analysis pipeline was required (Figure 3-2).

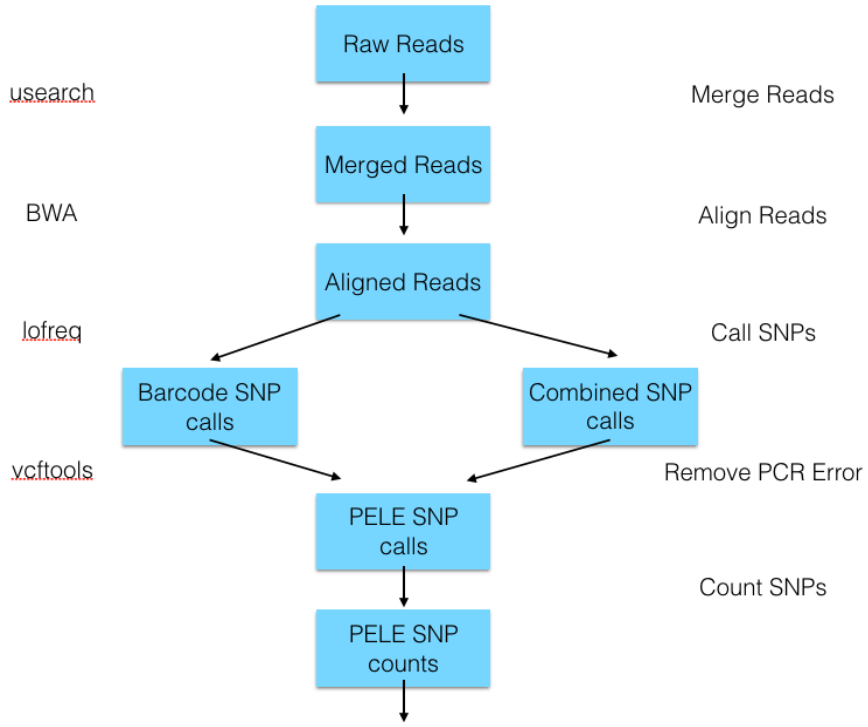

Figure 3-2. Analysis pipeline for PELE-Seq data.

To deal with overlapping reads, USEARCH was used (Edgar, 2010). It merges the forward and reverse reads and can trim the overhang, resulting in high confidence base calls. Reads are then aligned with BWA (Li, 2009) and SNPs are called with LoFreq (Wilm, 2012). Unlike most available SNP callers, LoFreq is not haplotype based and can call low frequency alleles with confidence. The barcode information was incorporated by calling SNPs on both barcodes together with more stringent filters and separately with more relaxed standards. All SNPs called in the combined sample were then verified in the individual barcode SNP calls with VCFtools (Danecek, 2011). Any SNP called in the

combined sample that was also present in both individual barcodes was considered a high-confidence SNP. This allowed us to have the greatest sensitivity with the fewest false positives.

To characterize mutations arising in tumorigenesis, tumor and non-tumor DNA from an ovarian cancer patient was purchased from Origene (CD564858 and CD564866, Table 3-1). Mitochondrial DNA was amplified using two sets of primers that produced fragments around 7500bp in length, which were then size selected to exclude remaining genomic DNA and made into Nextera sequencing libraries. This process was repeated for each region to control for PCR error.

Mathed Tumor and Wild-Type Samples

	Tumor	Wild-Type
Tissue of Origin	Ovary / Omentum Tumor	Myometrium
Pathology	Adenocarcinoma of ovary, clear cell, metastatic	Within normal limits
Stage	IIIC, moderately differentiated	–

Table 3-1. Matched tumor and non-tumor sample information. Physical description and pathology of tissue of matched tumor and non-tumor samples DNA obtained from Origene.

To look at the spatial distribution of mtDNA mutations, DNA from six sections of a single human brain tumor (Figure 3-3) was prepared in the same manner. This allows for detection of background mtDNA sequence (black circles), low level variants shared in all sections (blue circles), low level variants shared by a few sections (pink circles), and low level variants unique to a single section (green circle). Once sequencing was complete,

reads were processed as described in Figure 3-2 and variants were compared between tumor and non-tumor and among tumor sections to assess spatial distribution.

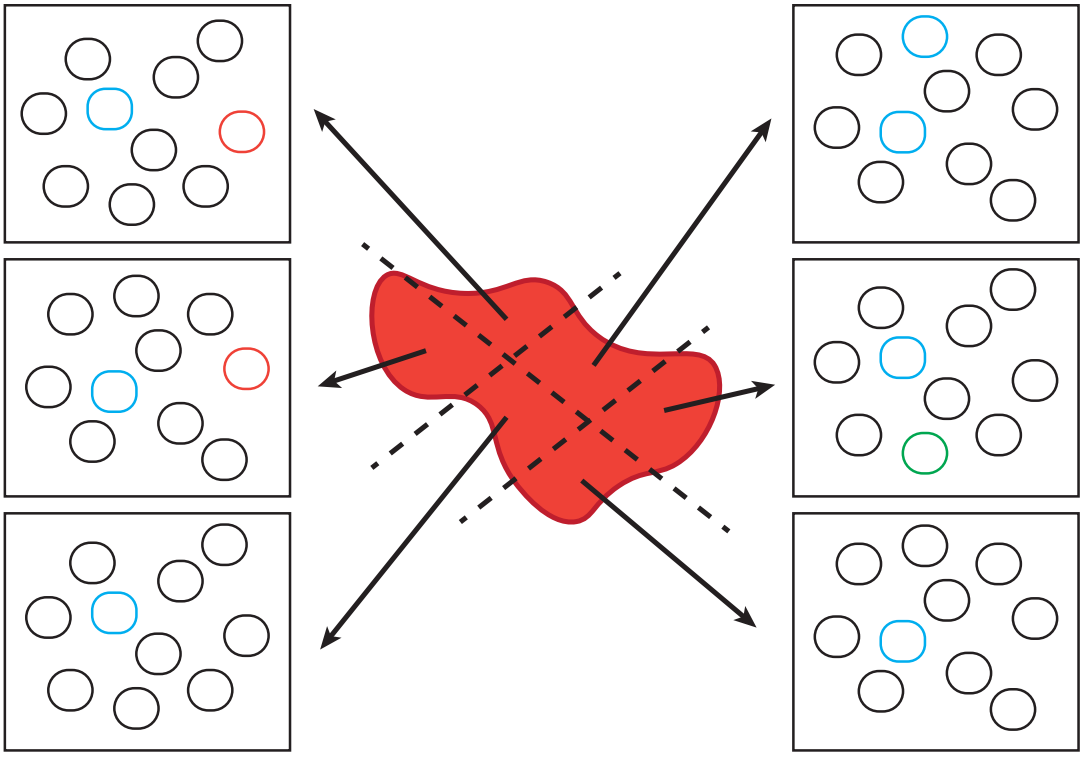

Figure 3-3. Sampling of Glioblastoma for Mitochondrial DNA Extraction. Solid tumor was sectioned into 6 pieces as shown to determine the spatial distribution of mitochondrial variants. A theorhetical pattern is shown for each of the sections. Blue circles represent ancestral mitochondrial haplotypes, pink represent new variants formed early in tumorigenesis and the green circle represents a relatively new variants formed late in tumorigenesis.

C. Results

SNP Detection in HEK cell line. We first wanted to test variant detection in a PCR amplicon of mitochondrial DNA from total DNA isolate. We were able to create PELE-Seq libraries from a mtDNA amplicon from total DNA isolated from a HEK293 cell line. We found SNPs ranging in frequency from 0.07 to 100% in the mitochondrial population (Table 3-2). SNPs covered the entire amplicon with at least 1000X coverage of most genes.

Human Embryonic Kidney Cell Line Heteroplasmy

Position	Alternate Allele	Coverage	Alternate Allele Percentage	dbSNP?	Genomic Region
750	G	1450	100	yes	12S-rRNA
908	T	2439	0.082	no	12S-rRNA
1123	T	1579	26.029	no	12S-rRNA
1236	T	2325	0.559	no	12S-rRNA
1590	G	2482	0.08	no	12S-rRNA
1644	T	2297	4.266	no	tRNA-val
2330	C	1081	0.185	yes	16S-rRNA
2645	T	1048	0.19	no	16S-rRNA
2706	G	654	99.84	yes	16S-rRNA
3107	T	1190	0.924	yes	16S-rRNA
3197	C	880	100	yes	16S-rRNA
3460	A	1187	0.168	yes	ND1
3836	T	2006	0.099	no	ND1
3955	A	1597	0.125	no	ND1
4235	A	1062	4.519	no	ND1
4464	T	1610	0.124	no	tRNA-met
4769	G	244	100	yes	ND2
5339	T	1739	100	yes	ND2
5347	T	2148	0.093	no	ND2
5982	T	2570	0.077	no	COI
6224	A	1850	0.108	no	COI
6462	C	2779	0.071	no	COI
7028	T	2011	100	yes	COI
7896	T	1508	0.331	no	COII

Table 3-2. Human Embryonic Kidney Cell Line Variants. For each allele at a given site, the total coverage, percent alternate allele, previous detection, and affected genomic region are shown.

SNP Detection in Paired Samples. Overlapping paired-end 150bp reads were down sampled to obtain 2M reads per PCR per amplicon, which amounted to roughly 7500X coverage. Many of the variants detected are shared between tumor and non-tumor samples (Figure 3-4 A). All homoplasmic alleles are shared as they represent the major haplotype of the individual from which both samples originated.

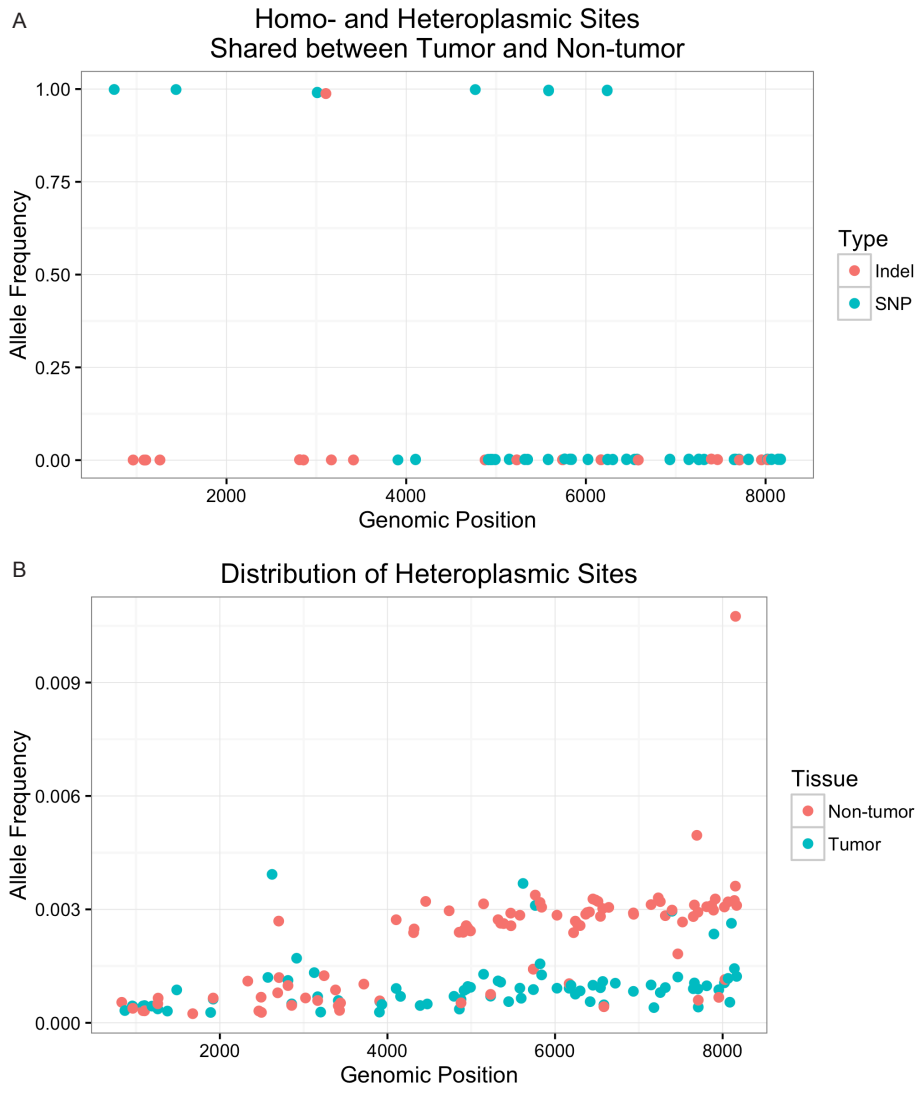

Figure 3-4. Distribution of Alleles in Matched Tumor and Non-Tumor Samples. (A) Distribution of indels (pink) and SNPs (blue) shared between tumor and non-tumor samples, across the mitochondrial genome. (B) Distribution of heteroplasmic sites across the mitochondrial genome, SNPs and indels in non-tumor tissue are in pink, SNPs and indels from tumor tissue are in blue.

There are more variants present in the 4-8kb region of the amplicon. Heteroplasmic sites are present at very low frequencies in both samples (<0.3%) which is consistent with previous work (Payne, 2013) (Figure 3-5 B). Non-tumor sample SNPa are at a higher frequency in the 4-8kb region than tumor SNPs. This pattern suggests common low frequency haplotypes in non-tumor tissue, from which a smaller pool formed the tumor population.

Figure 3-5. Changes in Allele Frequency of Shared SNPs. (A) Fold change of shared SNPs in tumor sample compared to non-tumor across the mitochondrial genome. Most shared alleles are less frequent in the tumor sample. (B) Distribution of fold changes for alleles in tumor sample compared to non-tumor. Indels are shown in pink and exhibit a normal distribution. SNPs are shown in blue and are lower in frequency in tumor samples.

This pattern is reinforced by the fold changes of SNPs shared between tumor and non-tumor tissue (Figure 3-5 A). All shared SNPs are at the same frequency in both tissues or are lower in frequency in the non-tumor tissue and variants in the 4-8kb region are almost all less frequent in tumor tissue. Shared indels do not follow the same trend and have a roughly normal distribution of fold changes while SNPs are skewed towards a low fold change (Figure 3-5 B). This implies either that the small population of progenitor cells to the tumor cells had fewer rare mtDNA haplotypes to begin with or that the rare mtDNA haplotypes confer some growth disadvantage and were selected against during tumorigenesis.

Looking at low frequency variants across the amplicon, an interesting pattern emerged. Those positions with shared low frequency alleles between the tumor and non-tumor samples show a marked physical distribution, arising only from 4-8kb and mostly absent from 0.75-4kb. Because there was a clear patter differentiating 0.75-4kb and 4-8kb, we wanted to determine the gene content of these regions (Figure 3-6). The major difference in these regions in the protein coding gene content. From 0.75-4kb, mtDNA primarily codes for rRNA and tRNA. From 4-8kb, there are several important protein coding genes (ND1,ND2, CO1, and CO2) which are involved in cellular respiration.

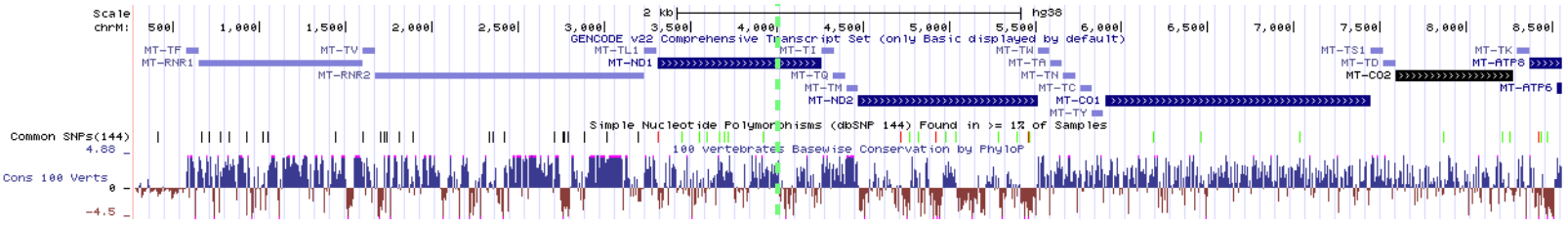

Figure 3-6. UCSC genome view of sequenced region of mitochondrial genome. Green vertical line correlates with lines in SNP and indel frequency plots at 4kb.

A similar pattern emerges when the variants unique to each sample are plotted along the amplicon (Figure 3-7) . There are differences in the gene content of these regions could account for the differences in variant distribution seen. The region from 750bp to 4kb harbors very little protein coding sequence, and is mostly made up of rRNA genes, which tend to be more permissive of variation, while the region from 4kb to 8kb contains much more protein coding sequence, which could be more sensitive to indels causing frameshift mutations.

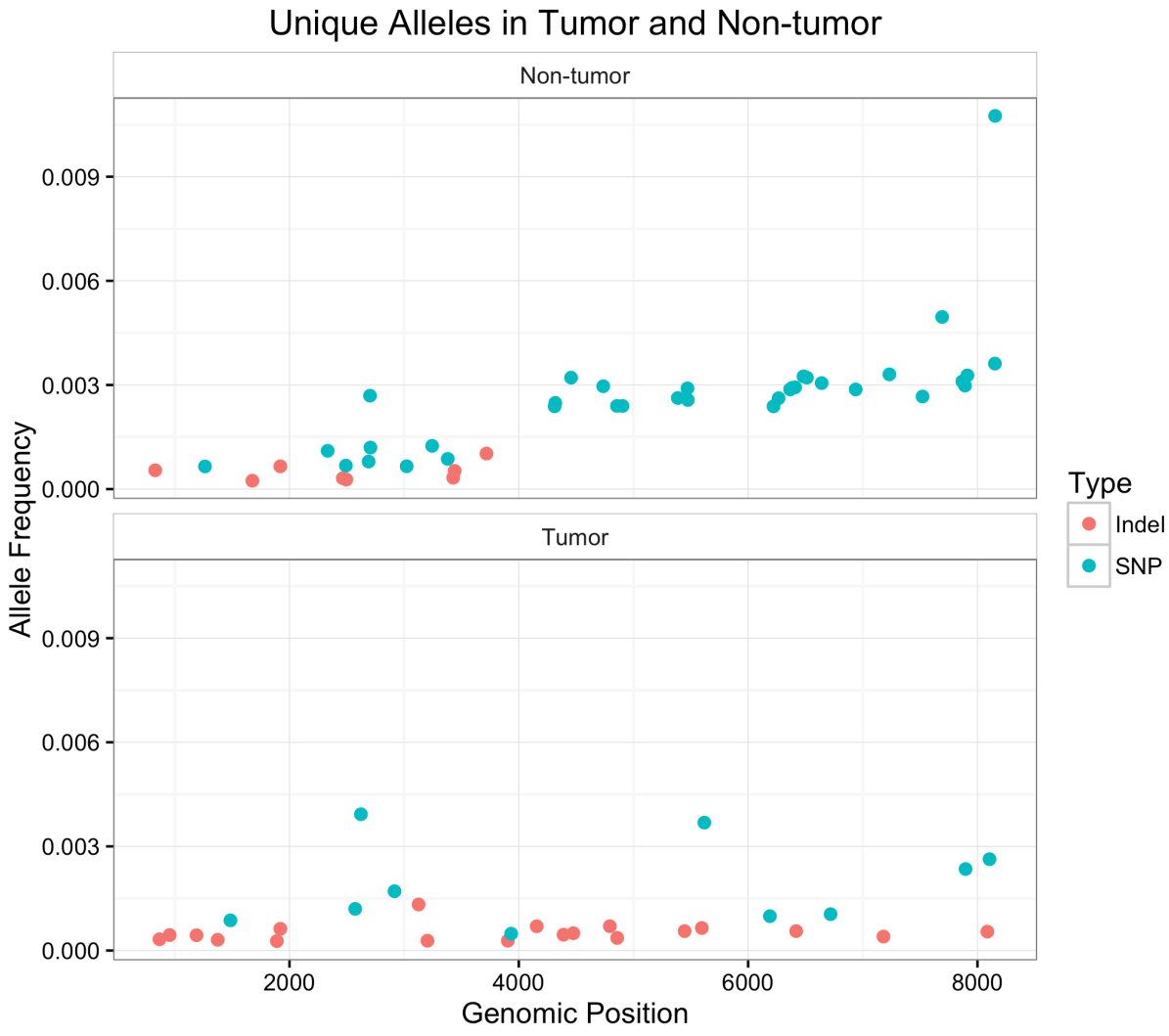

Figure 3-7. Variants Unique to Tumor and Non-tumor Samples. SNPs (blue) and indels (pink) unique to either non-tumor sample (top panel) or tumor sample (bottom panel).

Overall, the tumor sample has many more unique indels than the non-tumor sample, while the non-tumor sample has more unique SNPs. Fewer unique SNPs in the tumor sample suggests the tumor cells arose more recently from a smaller subpopulation of the non-tumor cells. The non-tumor sample has no indels in the 4kb to 8kb region which it does not share with the tumor sample, suggesting an environment more permissive of mutation or with different metabolic requirements.

We predicted the impact of the variants unique to the tumor sample using the Ensembl variant effect predictor (McLaren, 2016) and summarized the results in Table 3-3. While most are low impact, there are a few high impact frameshift variants in protein coding regions which could have selective consequences for these haplotypes.

Predicted Impact of Tumor Variants

Location	Allele	Consequence	IMPACT	GENE
4136	C	frameshift variant	HIGH	MT-ND1
4136-4137	-	frameshift variant	HIGH	MT-ND1
4958	G	synonymous variant	LOW	MT-ND2
5147	A	synonymous variant	LOW	MT-ND2
5147	A	regulatory region variant	MODIFIER	-
5320	T	missense variant	MODERATE	MT-ND2
5320	T	regulatory region variant	MODIFIER	-
5351	G	synonymous variant	LOW	MT-ND2
5351	G	regulatory region variant	MODIFIER	-
5742-5752	-	TF binding site variant	MODIFIER	-
5821	A	non coding transcript exon variant, non coding transcript variant	MODIFIER	MT-TC
		regulatory region variant	MODIFIER	-
5840	T	non coding transcript exon variant, non coding transcript variant	MODIFIER	MT-TY
		regulatory region variant	MODIFIER	-
6168	C	frameshift variant	HIGH	MT-CO1
		regulatory region variant	MODIFIER	-
6168	-	frameshift variant	HIGH	MT-CO1
		regulatory region variant	MODIFIER	-
6452	T	synonymous variant	LOW	MT-CO1
7663	T	synonymous variant	LOW	MT-CO2
8065	A	synonymous variant	LOW	MT-CO2

Table 3-3.Predicted Impact of Tumor Variants. Each variant given to Ensembl's VEP is shown with the type of variant it causes, the predicted level of impact, and the gene which it effects.

Variant Detection in Tumor Sections. Sequencing of mtDNA libraries resulted in ~4000X coverage for most sites across the amplicon. Merging of forward and reverse reads removed sequencing errors as expected. Analysis of SNPs yielded variants present at high frequency and in all regions of the tumor as well as low frequency alleles, both shared among and unique to samples (Figure 3-8).

A few interesting patterns emerged from this analysis. First, there were several consensus variants shared between all sections as the dominant or only allele (large, red, outer circles, Figure 3-8). This is to be expected since each region came from the same tumor, whose starting mtDNA had a few differences from the reference mitochondrial genome. Second, there were a few variants present at low levels in all tumor sections (large, red, outer circles, Figure 3-8). These suggest that there is indeed some standing heteroplasmy in the cells prior to tumorigenesis that is then carried on throughout the tumor as it grows. Thirdly, there are SNPs present in 2 (large dark blue outer circles, Figure 3-8) or 3 tumor sections (large, light blue, outer circles, Figure 3-8).

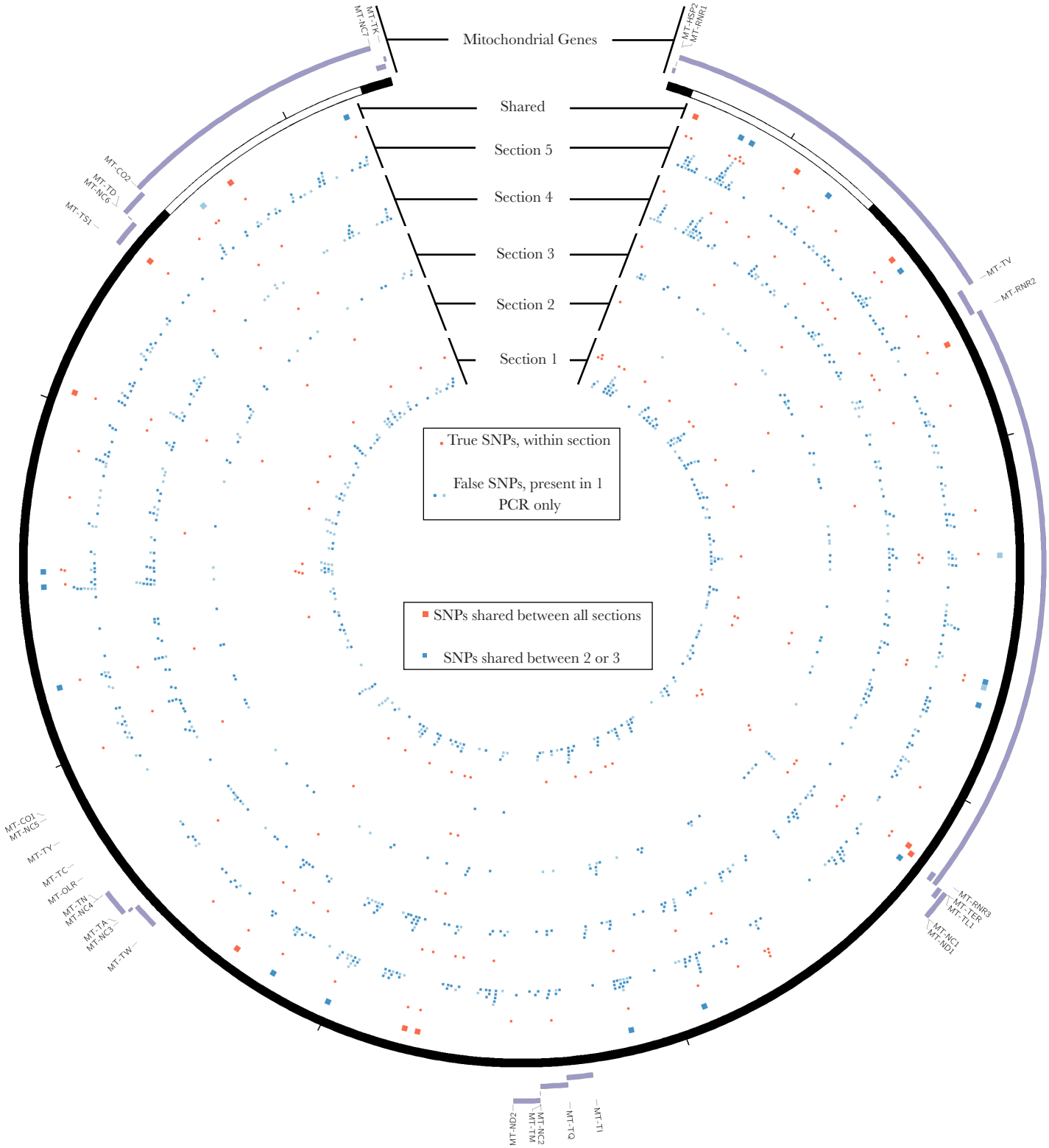

Figure 3-8. SNPs found in each section and in the tumor as a whole plotted along the mitochondrial amplicon. Inner blue rings show SNPs called in either the first PCR (light blue) or the second (dark blue), but not both. The red circles represent the true SNPs called in both PCRs. The outermost ring with larger dots represents the SNPs present in all tumor sections (red), 3 tumor sections (light blue) or 2 tumor sections (dark Blue). Mitochondrial genes are plotted outside in purple.

From this analysis, we were able to determine that variants are spatially distributed within the tumor (Figure 3-9). This suggests that mutations arise as the tumor forms, and are passed on to the progenitor tumor cells. The same mutations are not arising across the tumor, and so certain variants are absent from particular regions.

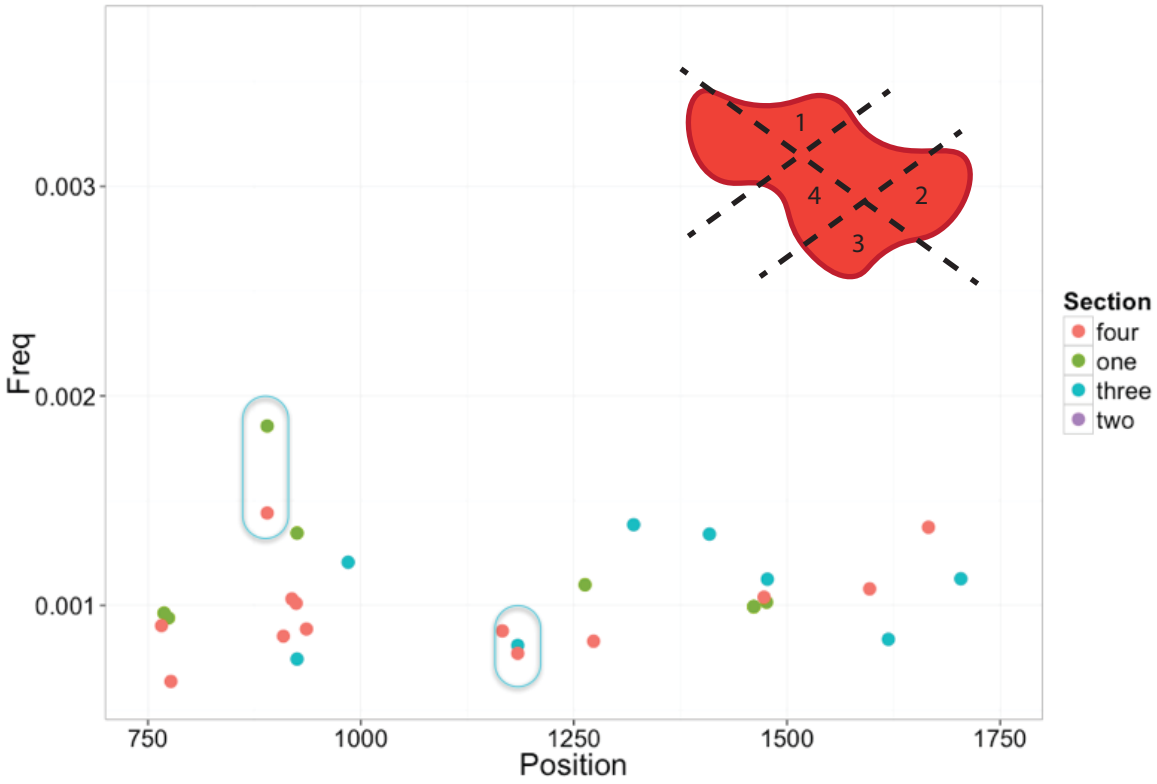

Figure 3-9. Variants are spatially distributed within the tumor. SNP frequency for each of the four sections is mapped to its genome position. Blue ovals indicate SNPs shared between two sections, but absent from the other two sections.

We also saw that the same variant was present in different sections at different frequencies (Figure 3-10). This is consistent with mutations arising as the tumor develops. Because each section is a large number of cells, different frequencies of the same allele in different sections indicates a different proportion of cells arising from the cell with the initial mutation.

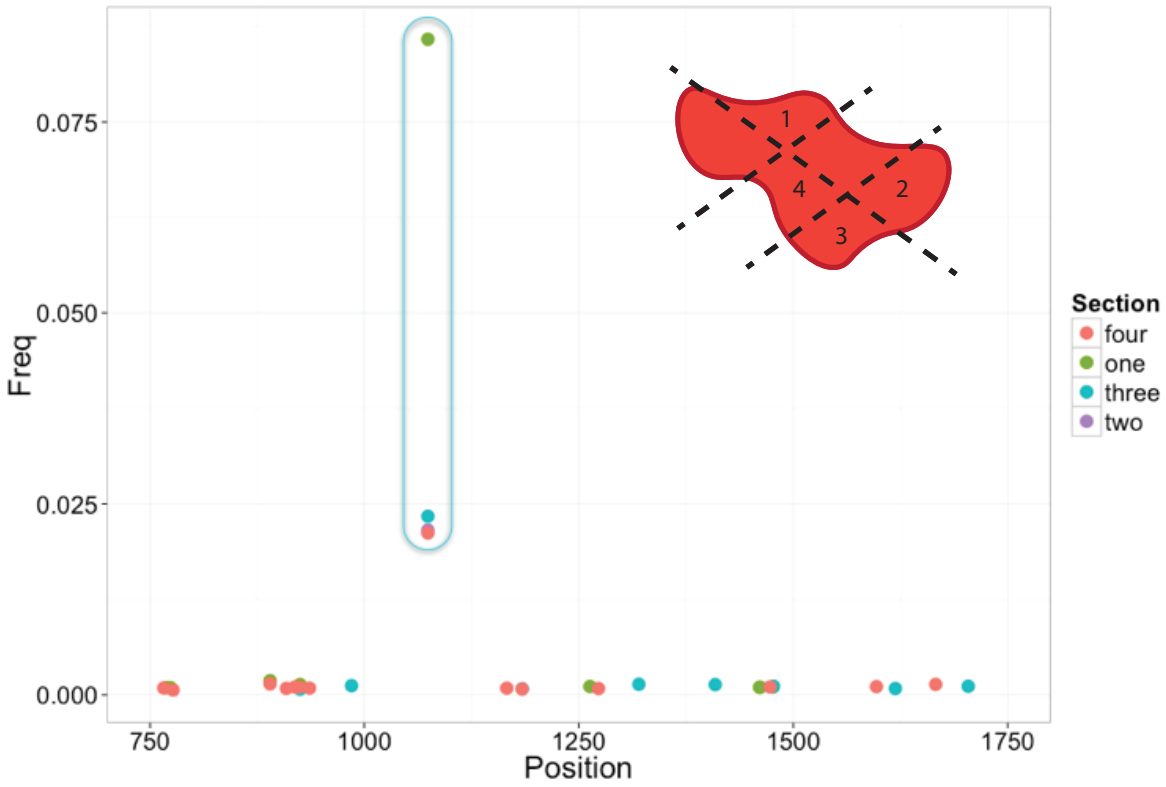

Figure 3-10. Variants are present at different frequencies in different sections of the tumor. SNP frequency for each of the four sections is mapped to its genome position. Blue oval indicates SNP shared among three sections, but present at different frequencies within those sections.

These patterns show that mutation is occurring as the tumor grows, making spatial patterns of mutations. Most mutations are unique to a single section of the tumor suggesting that new mutations accumulate quickly and throughout the tumor growth process.

D. Conclusions

PELE-Seq has been applied to sequence rare mutations in tumor mitochondrial genomes and matched tumor and wild-type DNA from an ovarian cancer mitochondria has been analyzed. With our new, sensitive detection of rare variants we were able to see distinct patterns of variation in different regions of the mitochondrial genome in tumor and non-tumor cells from the same subject, potentially because of different selective pressures on each region. We were also able to detect spatial organization in different regions of a solid tumor. Because most of the variants in the tumor are unique to a specific section, it appears that the overall mutation rate in the tumor is high.

BRIDGE

Once the experimental and computational pipeline for distinguishing rare variant from sequencing error and noise was used to interrogate variation in tumor mitochondria, we applied the PELE-Seq Approach to variation in somatic cells of zebrafish and mouse models for the DNA damage repair disease, Fanconi Anemia.

CHAPTER IV

SOMATIC MUTATION IN FANCONI ANEMIA

A. Introduction

Fanconi Anemia (FA) is an autosomal recessive inherited DNA damage disorder resulting from the loss of one of the fanconi anemia proteins. These proteins are involved in two main complexes that mediate DNA damage repair. A 2005 study in human B-lymphoblastoid cell lines used spontaneous mutation of a specific gene as a proxy for overall mutation rate and found FA to have a 30-fold higher mutation rate than wild-type (Araten 2005). They were unable to characterize specific mutation patterns and were limited to cell culture studies due to the limitations of their technique. Aside from large chromosomal rearrangements, the spectrum of somatic mutation in FA remains largely unexplored.

The full gamut of fanconi proteins is conserved from zebrafish to humans, making zebrafish an excellent model for this disease. By combining PELE-Seq with a longitudinal sampling scheme in fanconi anemia model zebrafish and multiple tissues from zebrafish and mice, it should be possible to elucidate the fine scale mutation patterns resulting from this disease, in growth, regrowth, specific tissues, and across species.

B. Methods

To examine the mutation rate of fanconi anemia model zebrafish, a cross was set up as shown in Figure 4-1. A male and female fish heterozygous for a fancd1 mutation were crossed and the offspring were raised to 2 months old, when their caudal fins could be clipped. Six weeks later the dorsal fins were taken and six weeks after that the anal fins were clipped. The fish were allowed to recover and all three fins were taken again. The first 3 fin clips will allow for determination of mutation accumulation over time in non-

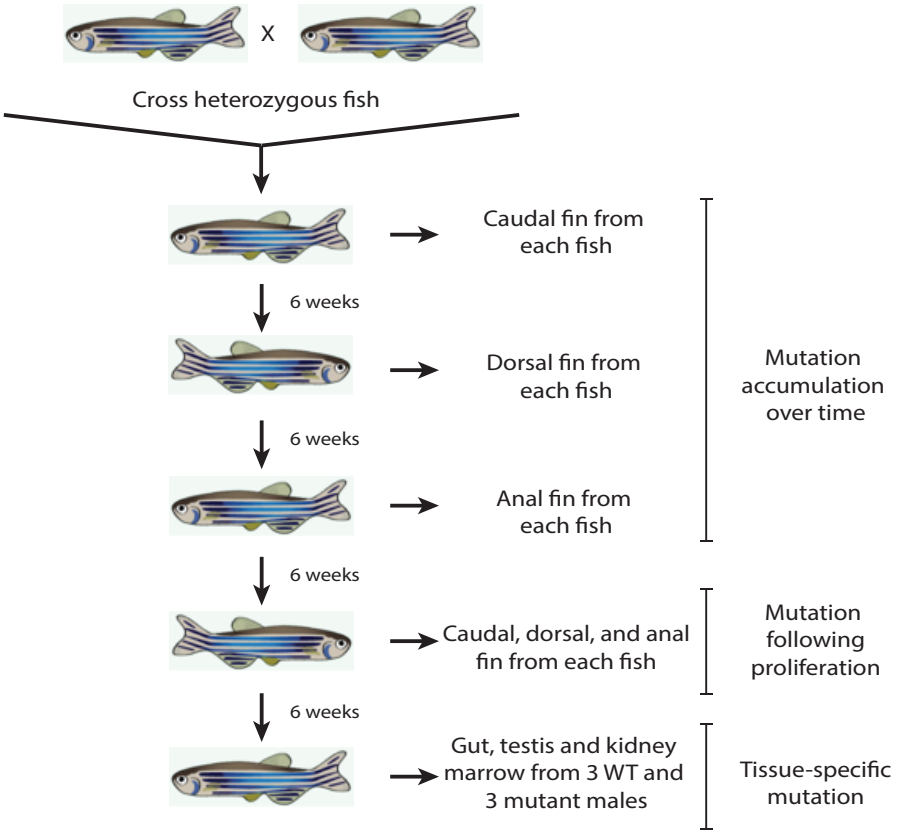

Figure 4-1. Sampling scheme for Fanconi Anemia zebrafish and their heterozygous and wild-type siblings. Offspring were obtained from a single heterozygote cross. Once the fish reached adulthood, a different fin was clipped every six weeks to examine mutation accumulation over time. Then all three main fins were clipped to look at mutation accumulation following proliferation. Finally, the gut, testis, and kidney marrow was harvested from 3 mutatnt and 3 wild-type males to look at tissue specific muations.

regenerating tissue. The final 3 clips will show mutations that arose following the regeneration of each fin. Finally, the gut, testis, and kidney marrow were taken from 3 wild-type and 3 mutant male fish to investigate tissue specific mutation.

Because of the large number of samples in this experiment and the depth of sequencing required for PELE-Seq, double-digest RADseq (ddRADseq) will be employed to sample about 40,000 bases of the genome(Figure 4-2 A) (Petersen 20121).

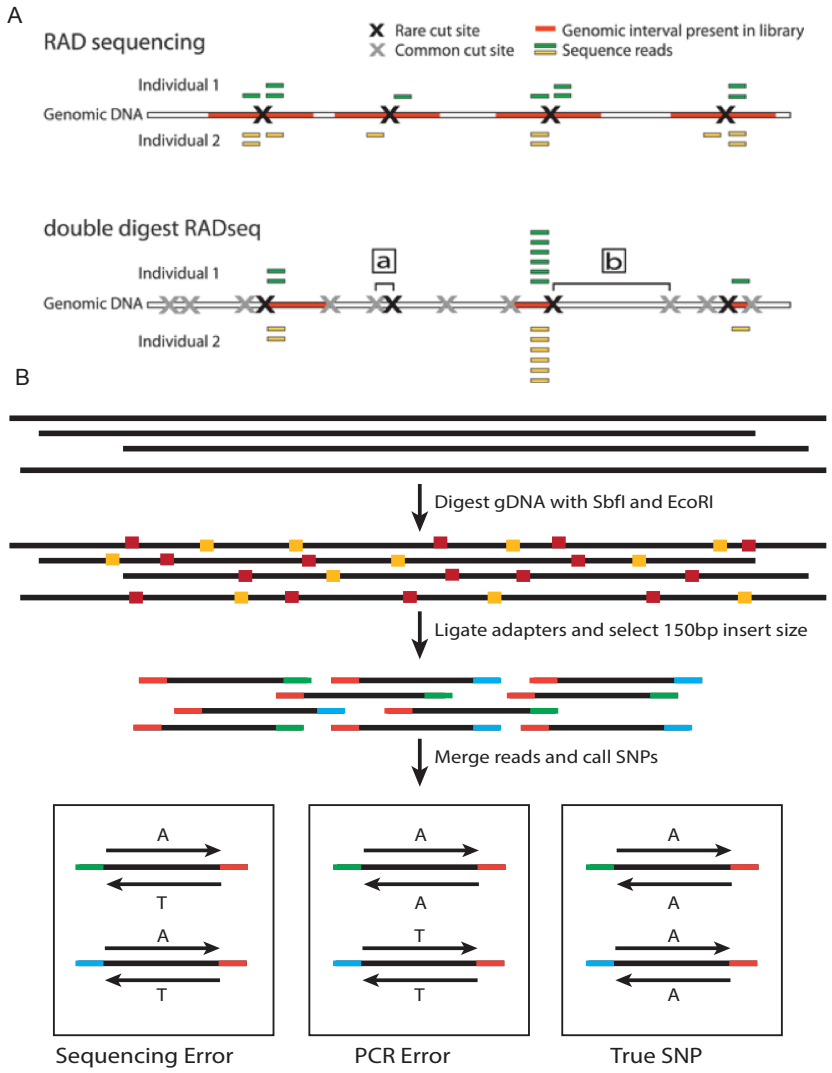

Figure 4-2. Experimental Method for ddRAD PELE-Seq. (A) RAD-Seq vs ddRAD-Seq, adapted from Peterson, *et al*, 2012. Instead of a single restriction enzyme, two restictions enzymes are used to digest DNA to allow for more specificity. (B) For this experiment, zebrafish gDNA was digested with SbfI and EcoRI and adapters were ligated. Inserts of 150bp were selected and the library was sequenced on a paired-end 150 Illumina HiSeq 2500 lane.

ddRADseq works very similarly to RADseq, but uses 2 restriction enzymes and a size selection step rather than one enzyme and random sampling. Because a consistent set of fragments is generated between restriction sites, size selection can be used to downsample the number of sites consistently between samples. An average SbfI RADseq library contains upwards of 100,000 RAD sites. By combining the restriction enzymes SbfI and EcoRI and size selecting fragments containing 150bp inserts (necessary for PELE-Seq) this is reduced to <500 sites (Figure 4-2 B). For each sample, at each time point, a set of ~200 one hundred base pair tags were be generated which will then be compared for mutation occurrence. DNA samples from the fins of 8 wild-type, 8 mutant, and 8 heterozygote fish were obtained as in Figure 4-1, as well as gut, testis, and kidney marrow tissue from 3 wild-type and 3 mutant fish. The DNA was prepared into an SbfI-EcoRI ddRAD library and size selected for 100-150 bp inserts.

To look at mutation patterns induced by fanconi anemia mutants across species, mouse samples were obtained according to Figure 4-3.

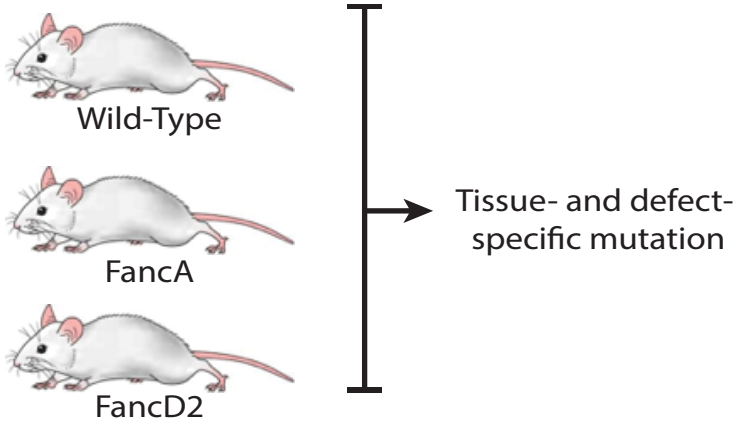

Figure 4-3. Sampling scheme for mouse Fanconi Anemia samples. DNA was extracted from liver and spleen from wild-type, FancA mutants, and FancD2 mutants.

These samples allowed comparison of different fanconi mutants within a species (fancA and fancD2 mutants), across different tissues of the same mutant (liver and spleen tissue). Reads were generated for both barcodes of each of the samples outlined in Figures 4-1 and 4-3. Reads were processed as described in Figure 3-2. Aligned reads in combined samples were used to define loci where all samples had at least 400X coverage. These loci (about 150 for zebrafish) were used to restrict reads to only these regions and downsample the high coverage bam files to even coverage across all samples. Realignment and indel quality assignment were performed with LoFreq, which was then also used to call SNPs and indels for each of the barcode sets and the combined read files. Per sample frequency for these sites was taken from the combined read calls.

C. Results

This analysis was first applied to the tissue samples from WT and Fancd1$^{-/-}$ zebrafish, as this subset had the best coverage and sample size to develop an analysis pipeline. Principal component analysis using the frequency of each detected variant in each sample was performed in R and samples were clustered (Figure 2-4). Samples group loosely by genotype (Panel A) and tightly by fish (Panel B). This suggests there is a slight genotype-dependent effect on the number or frequency of variants in each fish.

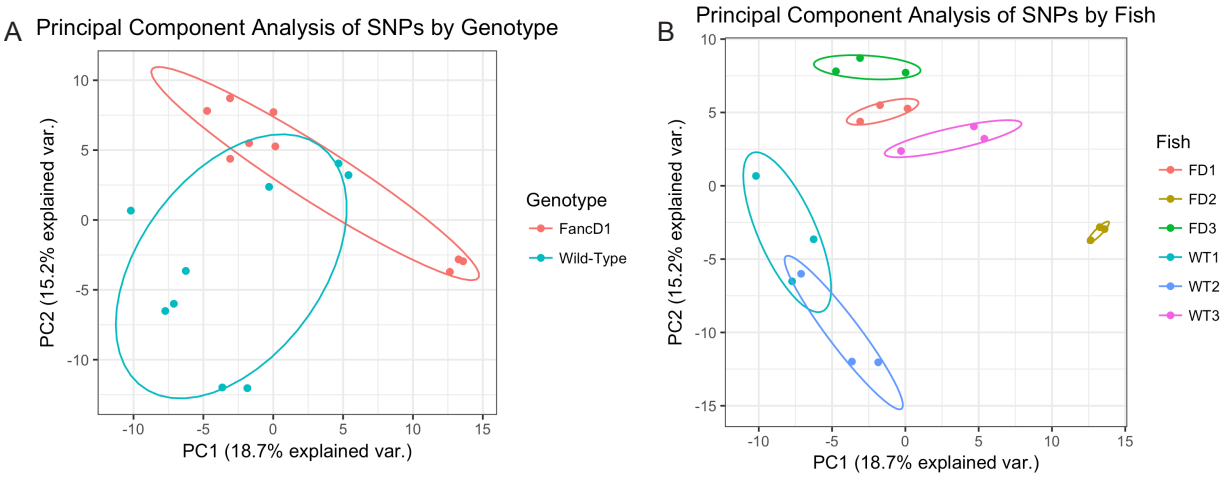

Figure 4-4. Principal component analysis and clustering of zebrafish tissue samples. (A) PCA plot grouped by genotype. Samples loosely group by genotype (B) Same PCA plot grouped by fish. Samples group well by fish.

We next looked at site GC content and length to see if these were affecting variant distribution (Figure 4-5). Sites were divided into four groups: sites containing no variants, sites with equal numbers of variants in both genotypes, sites with more variants in FancD1 mutants, and sites with more variants in wild-type fish. There was no difference in distribution for site length or GC content for any of these groups.

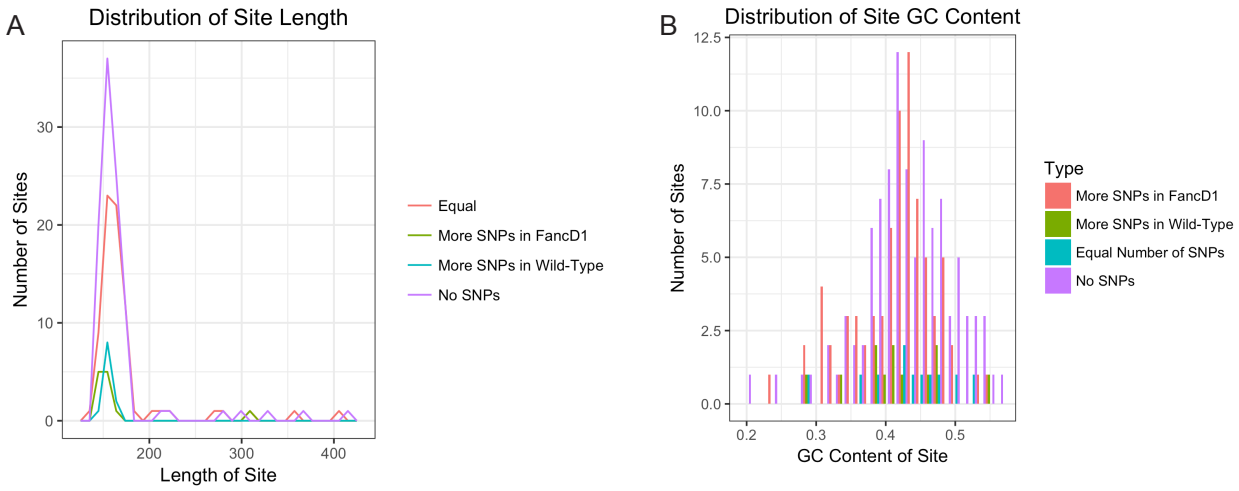

Figure 4-5. Site length and GC content distribution across zebrafish tissue samples. (A) Distribution of site length centers tightly around 150bp for all types of site: those with equal number of variants in FancD1 and Wld-Type fish (pink), those with more variants in FancD1 fish (green), those with more variants in Wild-Type fish (blue), and those with no variants in either genotype (purple). (B) Distribution of GC content of sites, divided as in panel A. In all cases, GC content centers around 0.5.

Next, the number of SNPs and indels per individual for each genotype was examined (Figure 4-6). Wild-type and FancD1 fish did not have any difference in the total number of SNPs per individual, but there were slightly more indels per individual in FancD1 fish.

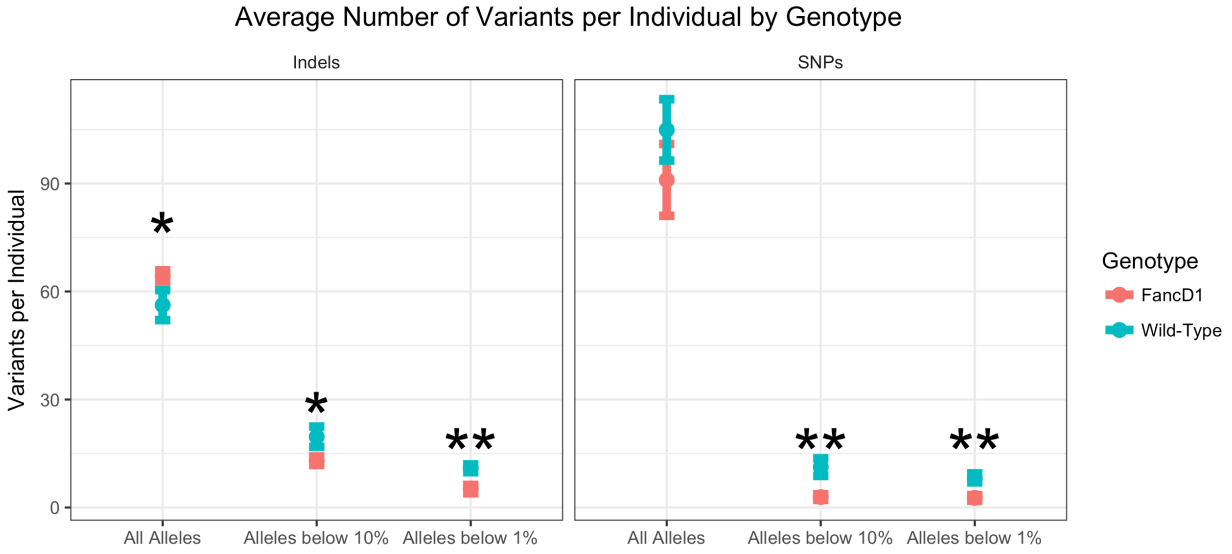

Figure 4-6. Variants per individual at different frequency cutoffs show significant differences in number of variants in wild-type and fandD1 fish. Average indel count per individual shown on left for all alleles, alleles below 10%, and alleles below 1%. Average SNP count per individual for the same cutoffs is shown on the right. Averages for FancD1 individuals (pink) and wild-type individuals (blue) are shown for each group. Stars indicate significance: * $p < 0.1$, ** $p < 0.01$.

Low frequency alleles (both SNPs and indels below 10% and below 1%) were significantly less frequent in FancD1 fish. This could be explained by faulty error-prone DNA repair pathways in FancD1 mutant fish. Without the ability to read through stable replication forks, FancD1 mutant cells must either resort to more drastic repair measures such as double stranded break repair or stall in growth and enter apoptosis. Perhaps due to smaller sample sizes when considering tissue type, there were no significant differences in or between tissue types for SNP or indel count per individual (Figure 4-7).

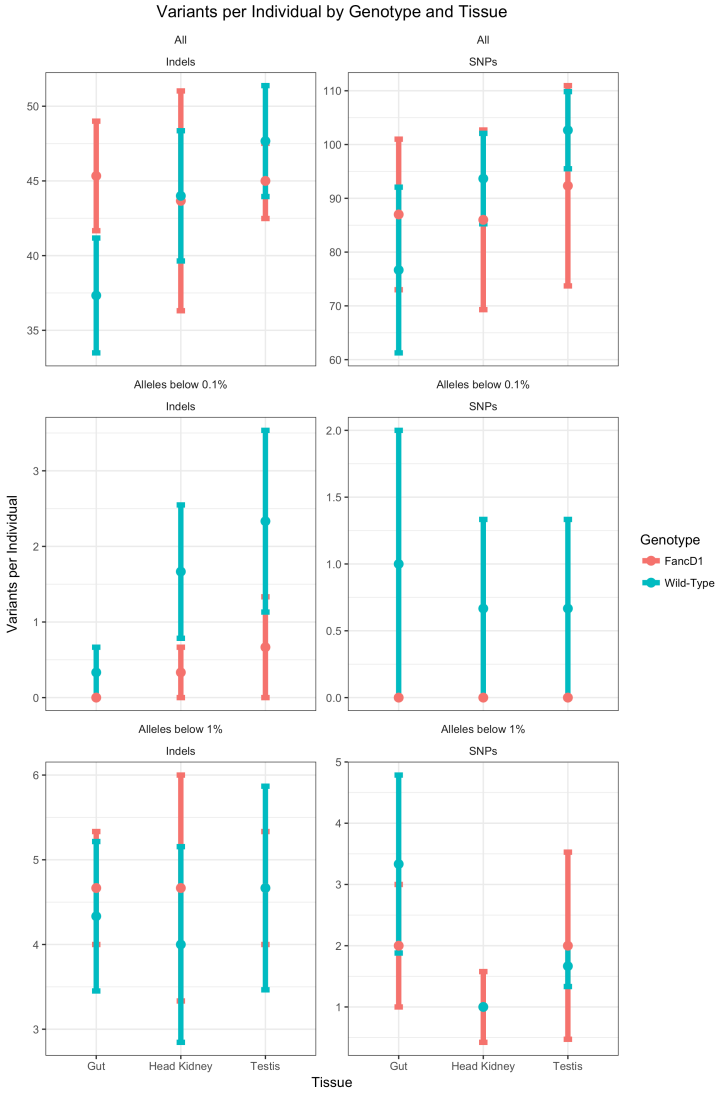

Figure 4-7. Variants per tissue per individual do not show genotype dependent effect. Average indel count per individual shown on left for all alleles, alleles below 10%, and alleles below 1%. Average SNP count per individual for the same cutoffs is shown on the right. Averages for FancD1 individuals (pink) and wild-type individuals (blue) are shown for each group.

Next, we wanted to look more closely at the types of mutations occurring. The transition-transvehrsion ration (Ts/Tv) was determined for all SNPs in each sample as well as low frequency SNPs (Figure 4-8). As in the overall numbers of SNPS, the Ts/Tv is not significantly different for all SNPs between WT and mutant tissues, but there is a difference in the low frequency SNPs.

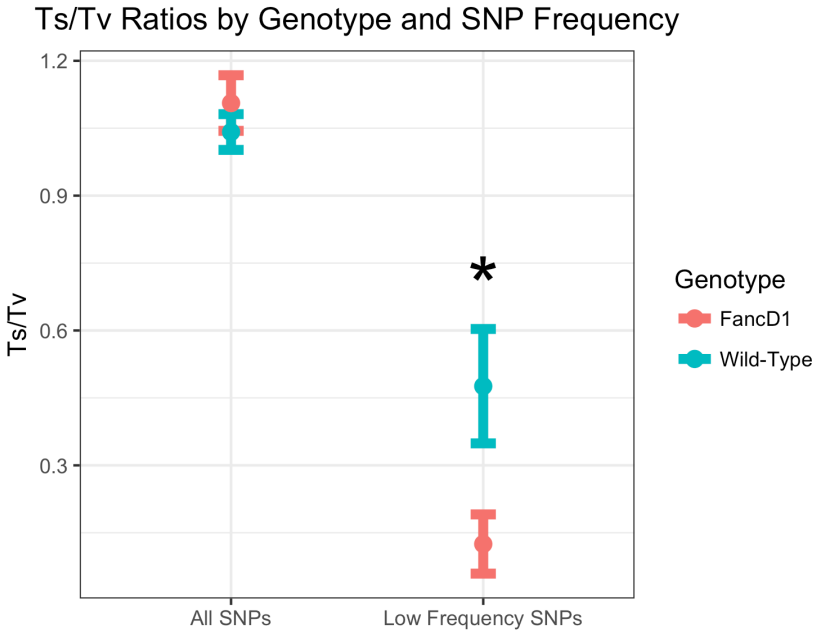

Figure 4-8. Ts/Tv ratio in All SNPs and low frequency (<10%) SNPs. FancD1 mutants (pink) have a significantly ($p < 0.1$) lower Ts/Tv ratio in Low frequency SNPs than wild-type (blue).

D. Conclusions

We have been able to create a library prep and analysis method to look at rare variation across many genomes. With this method, we have been able to determine that SNPs and small (<30bp) indels are more frequent in WT than in fanconi mutant tissues, potentially due to impaired error-prone polymerase recruitment in the mutant fish. Zebrafish fin tissue data did not show any discernible patterns, perhaps because, even after regeneration, too few cell division cycles had occurred to show a measurable effect.
In the future, larger structural variants will be investigated. It is known that FA affected individuals exhibit more structural variants than non-affected. Linked read technologies such at 10x genomic's platform allow for easy determination and phasing of structural variation and would be better suited to investigating variation in this system.

BRIDGE

We were able to use the experimental and computational pipeline for distinguishing rare variants within populations from sequencing error and noise in two heterogeneous populations. This still did not address sequencing rare members of complex communities and so we developed a method for enriching rare members of a complex biological association.

CHAPTER V

EFFICIENT TRANSCRIPTOME PROFILING OF HOST-ASSOCIATED BACTERIA

A. Introduction

Vertebrate gut-associated bacteria have been shown to play essential roles in the health and development of their animal hosts, including facilitating digestion and nutrient acquisition, education and maturation of the immune system, and protection from pathogens (reviewed in Neish, 2009). Our understanding of these roles has been transformed by sequencing technologies that allow an unbiased look at the composition and activity of bacterial communities and the development of model animal systems for mechanistic studies into these intimate biological relationships. Traditional approaches to understanding bacterial communities such as 16S sequencing provide taxonomic data, but do not assess total gene content of the community or the genome-wide expression data. Transcriptomics provides information about both gene content and the relative activity of the genes within and across conditions. Previous work has shown the intestinal environment to be highly dynamic and that the spatial structuring of different bacterial species within the gut undergoes dynamic responses to the changing environment (Wiles 2016). Tying transcriptional changes to specific phenotypes enriches out understanding of the genetic underpinnings of those phenotypes and gives new insight on ways to manipulate host-associated microbes for specific goals.

Transcriptomic data from many host-microbes system is often difficult to generate, largely because bacterial transcripts are rare (<0.1% of total RNA in larval zebrafish

guts). To make sequencing the bacterial transcriptome feasible, the proportion of bacterial RNA in the sequencing library needs to be increased. Different approaches are used to accomplish this, including host removal and hybridization capture methods. Host removal techniques attempt to eliminate the dominant, non-target species by selective hybridization and removal (Kumar 2016). However, host removal techniques are less effective when the ratio of host:microbial RNA is too large. In contrast, hybridization capture methods enrich for the minor species by binding the target RNA and removing the unbound, non-target RNA (Carpenter 2013). Hybridization capture has been used to effectively enrich very minor components within a mixed sample.

For this work, germ-free larval zebrafish were mono-associated with a bacterial strain, which was isolated from adult zebrafish guts, and is closely related to *Vibrio cholerae*. Zebrafish are an ideal system as they can be derived germ-free, then inoculated with a defined community. Indeed, zebrafish is an established model for intestinal community studies with a growing body of knowledge about the interactions between the host and commensal microbes (Stephens 2015, Wiles 2016, Rolig 2015, Hill 2016). The *Vibrio* isolate used in our study is a robust colonizer of larval zebrafish, providing a good target bacterium for testing enrichment from whole gut samples.

Our method is adapted from the whole-genome capture method described by Carpenter *et.al.*. who used extant human gDNA to create biotinylated RNA probes to enrich for human DNA from ancient bones and teeth. Their pre-enrichment libraries consisted of ~1.2% human DNA; post-enrichment, the majority mapped to the human genome. Because our system contained similar starting ratios (0.1% bacterial, 99.9% zebrafish),

we tested if a similar approach would allow us to generate bacterial transcriptomes from host-associated total RNA samples. Here,*Vibrio* gDNA was used as a template to generate biotinylated RNA probes. Standard RNA-Seq libraries were made from our experimental samples and then hybridized to the biotinylated RNA probes (Figure 5-1). We first used biological replicates of *in vitro* cultures of *Vibrio* to assess the efficiency and bias of our hybridization capture. We then compared transcription profiles of *Vibrio* associated with larval zebrafish hosts at 24 and 72 hours post-inoculation.

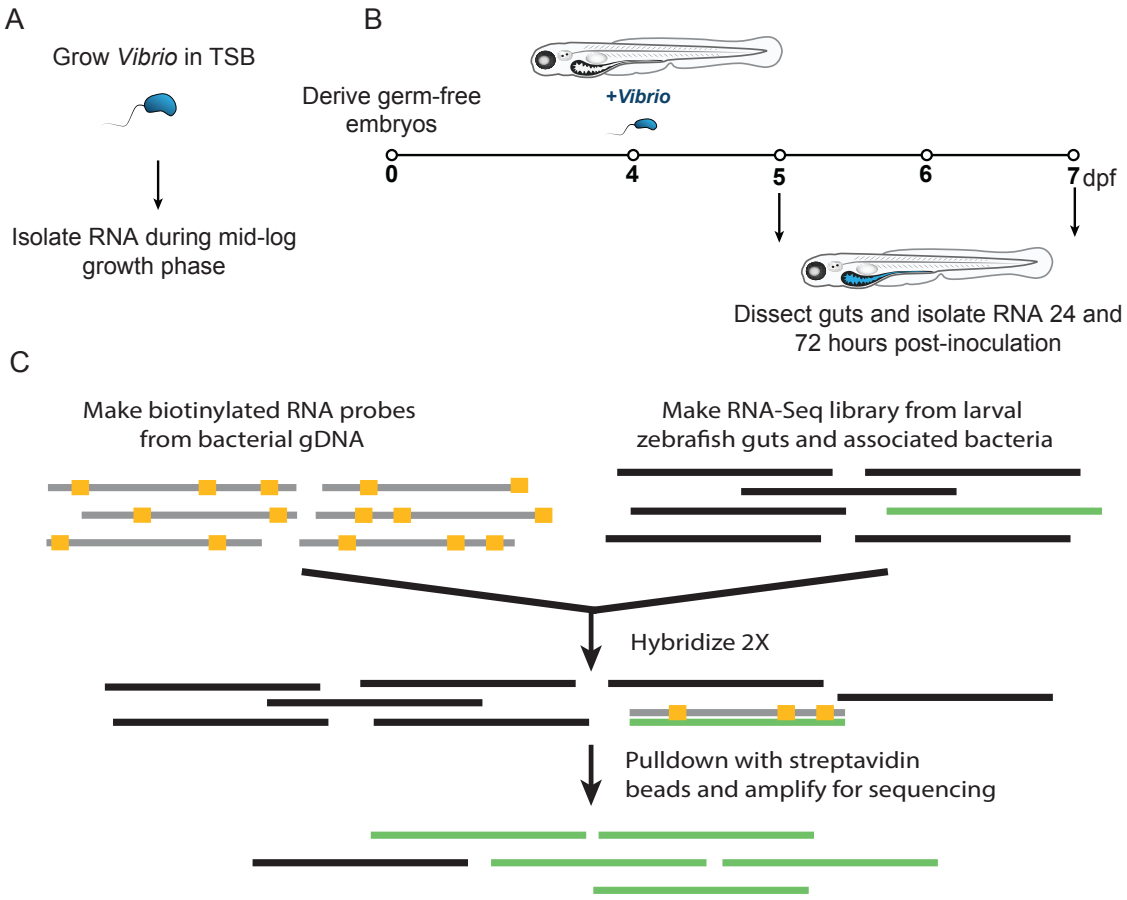

Figure 5-1. Experimental setup for validation samples and capture-hybridization method (A) For capture tests, biological replicate cultures of Vibrio were grown in TSB and RNA was isolated during exponential growth. (B) Sampling scheme for larval zebrafish. Germ-free embryos are derived and colonized 4 dpf with ZWU0020. Guts are dissected a 5dpf (24 hours post-inoculation) and 7 dpf (72 hours post-inoculation). Each sample in A and B were taken in triplicate. (C) Whole genome capture of bacterial transcriptomes. RNA-Seq libraries are made from total RNA extracted from zebrafish guts and associated microbes, which are then hybridized to biotinylated probes and amplified for sequencing.

B. Materials and Methods

Creation of biotinylated RNA probes. Probes were created from genomic DNA of the microbe of interest (*Vibrio cholera*) by fragmenting the genome into 300-800bp fragments, adding a T7 site, depleting rRNA sequences, and transcribing the library with biotinylated UTP to create biotinylated RNA probe libraries that were then used to capture bacteria RNA sequences. This allowed us to select for the bacterial RNA of our species while excluding the more abundant eukaryotic RNA. This also minimized the bias in selection of bacterial genes, since the entire genome was used to make the probe set.

To fragment the DNA, 2uL of high quality genomic DNA at 2.5 ng/uL was added to 2.5 uL TD Buffer and 0.5uL TD Enzyme from the Illumina Nextera kit (Illumina: FC-121-1031, FC-121-1012). This was incubated at 55°C for 5 min, then cooled to 10°C. To 4uL of this reaction, 1 uL PCR Primer Cocktail, 1uL each of 2 index primers, and 3uL Nextera PCR Master mix was added and the mix was amplified 10 cycles according to the Nextera amplification parameters. Reactions were cleaned using 1.8X MagBind Beads and eluted in TE. Ribosomal sequences were depleted using RiboMinus (Life-tech: K1550-04). 25uL Magnetic beads were washed twice in 25 uL RNase-free water and once in 25 uL Hybridization buffer (B10), then resuspended in 10 uL Hybridization buffer. In a separate tube, 18uL of the DNA library was mixed with 1 uL each P1 and P2 blockers (10uM DNA oligos), 4uL 1:10 dilution of RioMinus Probe from kit, and 25uL Hybridization Buffer and the incubated at 37°C for 5 minutes. Sample was immediately placed on ice for at least 30 seconds, then mixed with cleaned beads. This

was incubated at 37°C for 5 minutes, mixing occasionally. Beads were placed on magnetic stand and the ribo-depleted supernatant was collected and cleaned up with 1.8X MagBind beads. The T7 site was added by amplifying the 1.5 ng of the ribo-depleted DNA library with 1uL each of 10uM custom T7 site forward and reverse Nextera primers, 12.5 uL 2X Phusion master mix, and water to 25uL total. This was amplified for 8 cycles of 98°C for 10 sec, 60°C for 30 sec, and 72°C for 3 min. This was cleaned up with 1.8X MagBind beads. Finally, biotinylated RNA was synthesized from these templates using MEGAscript (Life-Tech: AM1334). Water to 20 uL, 10 uL DNA template, 2 uL 10X transcription buffer, 1 uL each 10 mM ATP, GTP, and UTP, 0.6 uL 10 mM CTP, 5 uL Biotin-16-dCTP (Life-tech: AM8452), and 2 uL T7 Enzyme mix were incubated at 37°C overnight (~16hrs). To remove template, 2uL Turbo DNase Buffer and 1uL Turbo DNase (Thermo-Fisher: AM1907) were added and incubated at 37°C for 20-30 minutes. Biotinylated RNA was cleaned up with Qiagen RNeasy columns (Qiagen: 74104), quantified by Qubit and stored at -80°C until capture libraries were ready.

Isolation of Vibrio RNA from *in vitro* and *in vivo* (larval zebrafish) samples. The bacterial strain used for this study was previously isolated from an adult wild type zebrafish, and designated ZWU0020. To generate in vitro RNA for ZWU0020, it was grown overnight in TSB (tryptic soy broth; BD worldwide, #211825), then back diluted 1:200 (10 ml total; three replicate cultures) into sterile TSB, and grown for 5 hrs at 30C, with shaking, to approximately mid-log growth phase. Three milliliters of each culture was mixed with an equal volume of ice cold methanol and iced for 15 min. Samples

were centrifuged to pellet cells, the supernatant was removed, and the cell pellets were resuspended in 200ul of RNAprotect (reagent (Qiagen), then placed at -80C until RNA extraction. RNA was isolated using the Zymo Quick-RNA miniprep kit (Zymo Research, #R1054), following product protocol, and eluting RNA in 30ul RNase-free water.

Zebrafish colonizations: Wild type zebrafish embryos were collected immediately after laying and made germ-free following the previously described protocol (Bates, 2006), and distributed into tissue culture flasks containing 15 ml embryo medium at a density of 15 embryos/flask. At 4 dpf (days post fertilization), 6 flasks were inoculated with 5ul of an overnight culture of Vibrio ZWU0020 grown in TSB medium, and washed once with embryo medium. At 5 dpf (24 hr colonization) and 7 dpf (72 hr colonization), 10 fish guts from each of three of the flasks were dissected, and the guts combined into 200ul of RNAprotect reagent, resulting in biological triplicate samples for each time point. For two of the 24 hr flasks, a single fish gut was dissected and placed into 200ul of RNAprotect reagent. 100ul of bullet blending beads (product) were added to each sample and they were bullet blended (machine info) for 1 min at power 4 setting. Samples were stored at -80 C until RNA extraction. RNA was extracted as described above for in vitro RNA samples. A schematic of the sample collection process is described in Figure 3-1A and 3-1B.

Synthesis of total RNA libraries. cDNA was reverse transcribed from the total RNA extracted from Vibrio-colonized zebrafish guts and *in vitro Vibrio* samples. Nextera libraries were made from the cDNA, pooled, and hybridized to the biotinylated RNA

probe library. These RNA/DNA hybrids were isolated with magnetic streptavidin beads, the RNA was destroyed, and the released enriched library was amplified and sequenced. To obtain host-associated RNA, larval zebrafish were mono associated with *Vibrio cholera*. Guts were dissected from the fish, pooled (~10 guts per pool), resuspended in 90uL TE and 10 uL Proteinase K and incubated at room temperature for 5 minutes. RNA was isolated with Qiagen RNeasy column according to the blood and tissue protocol and eluted in 30uL RNase-free water. Samples were DNase treated with Turbo DNase as before. After incubation, 4uL DNase inactivation reagent was added and incubated at room temperature for 5 minutes. Samples were centrifuged at 10,000xg for 1.5 minutes to precipitate inactivation reagent and the supernatant containing DNA-free RNA was saved. Using the Ribo-Minus kit, 125uL of magnetic beads were washed twice in 125uL RNase-free water, and once in 125uL Hybridization buffer, then resuspended in 40uL Hybridization buffer. In a separate tube, 15uL RNA, 2uL RiboMinus Probe, and 45uL Hybridization buffer were mixed and incubated at 37°C for 5 minutes, the remaining RNA was stored at -80°C. Hybridization mixes were incubated at 37°C for 5 minutes, then placed immediately on ice for at least one minute. Then the hybridized probes were mixed with the beads and incubated at 37°C for 15 minutes, occasionally mixing. Beads were placed on magnetic stand and the ribo-depleted supernatant was collected and cleaned up with the Qiagen RNeasy columns.

The Ovation RNA-Seq System V2 (NuGEN: 7102-08) was used to reverse transcribe and amplify cDNA using 5uL of each sample and following the manufacturer's protocol. cDNA was quantified and diluted to 2.5ng/uL for use in Nextera library prep. For each

sample, 4uL of 2.5ng/uL cDNA, 5uL Tagment DNA Buffer and 1uL Tagment DNA Enzyme were used to tagment the cDNA for 5 minutes at 55°C. From this reaction, 8 uL was combined with 2 uL PCR primer cocktail (PPC), 2 uL each one 50x and one 70x index, and 6 uL Nextera PCR master mix (NPM). Samples were amplified 8 cycles according to the Nextera amplification parameters. Samples were cleaned up using 1.8X MagBind beads and quantified individually. They were then pooled in equal parts by weight in preparation for hybridization and capture.

Enrichment of microbial RNA-Seq libraries. A schematic of the hybridization and capture protocol is presented in Figure 1C. The capture protocol was adapted from Carpenter, et al. Three mixes were prepared in PCR tubes for the hybridization: a DNA pond mix containing cDNA library pools and blocking DNA, a hybridization mixture containing the buffers and salts for the hybridization, and a bait mix containing the biotinylated RNA baits and RNaseOUT. The DNA pond mix contained 2.5 uL Salmon Sperm DNA (10mg/mL), 2.5 uL CotI DNA (1mg/mL), 2 uL each P1, P2, P1 nextera, and P2 nextera blockers (100uM each), 19 uL DNA library pool containing at least 300ng total DNA. The hybridization mixture contained 0.2 uL Water, 20 uL Hyb1 (20X SSPE), 0.8 uL Hyb2 (500 mM EDTA), 8 uL Hyb3 (50 X Denhardt's Solution), and 8 uL Hyb4 (1% Sodium Dodecyl Sulfate). The bait mix contained biotinylated probes, for 1.5x weight with DNA library pool, RNase-free water to 12uL, and 1.5 uL RNaseOUT (40U/ uL). The DNA pond was heated in a thermal cycler to 95°C for 5 min, followed by 65°C for 5 min. When the heat block reached 65°C, the hybridization mixture was added.

When the DNA had been at 65°C for 2.5 min, the RNA bait mix was heated to 65°C for 2.5 min in a heat block. At this point, 26uL of the hybridization mix was added to the DNA pond mix followed by 8uL of the bait mix. The reaction was incubated at 65°C for 12-16 hours.

Dynabeads® MyOne Streptavidin C1 (Thermo-Fisher: 650.01) were used to isolate DNA hybridized to RNA baits. 50uL of streptavidin beads were transferred into a new 1.5mL tube and pelleted with magnetic particle stand. Supernatant was discarded, and beads were washed three times in Binding Buffer (1 M NaCl; 10 mM Tris−HCl, pH 7.5; 1 mM EDTA), then resuspended in 200uL Binding Buffer. Hybridization solution was added to the beads and incubated for 30 minutes at room temperature on a rotator. Beads were pelleted and resuspended in 500uL Wash Buffer 1 (1X SSC, 0.1% SDS) and incubated at room temperature for 15 minutes. Beads were pelleted and resuspended in 65°C Wash Buffer 2 (0.1X SSC, 0.1% SDS) and incubated for 15 minutes at 65°C. The wash with Wash Buffer 2 at 65°C was repeated twice for a total of 3 washes. Beads were pelleted and resuspended in 50uL elution buffer (0.1M NaOH, prepared fresh) and incubated at room temperature for 10 minutes. Beads were pelleted and the supernatant was transferred into a new 1.5mL tube containing 70uL Neutralization Buffer (3.75 ml 1 M Tris−HCl, pH 7). DNA was cleaned up using 1.8X MagBind Beads and eluted in 30uL. Enriched bacterial DNA was amplified in two 50uL reactions containing 25uL 2X Phusion Master Mix, 2uL P5,P7 primers, 10uL sample, and 13uL water. Reactions were denatured at 98°C for 3 minutes and amplified for 16 cycles (98°C for 30 seconds, 65°C for 15 seconds, 72°C for 30 seconds) with a final extension of 2 minutes at 72°C.

Reactions were pooled, cleaned up with 1.8X MagBind Beads, eluted in 20uL. 1uL was used to quantify DNA and remaining 19uL were used as input for a second hybridization and capture, performed exactly like the first. After the second capture and amplification, libraries were sequenced.

Data analysis of microbial RNA-Seq libraries. After sequencing and demultiplexing, samples were aligned to an in-house ZWU0020 reference genome (IMG ID 99400) using BWA with default settings (Li 2013). The reference genome was generated from PacBio long read sequences and comprised 4243396 bases in 6 contigs, 2 of which are known plasmids. Genome annotation predicts 3805 protein coding genes, which were used as genes for differential expression analysis. Gene counts were obtained using bedtools (Quinlan 2010), and then passed to DESeq2 for differential expression analysis (Love 2014). Hierarchical clustering was performed in R as well to determine sample clusters. Differential expression data was passed to Gage (Luo 2009) to determine significantly enriched KEGG pathology clusters, and the results were visualized using Pathview (Luo 2013). COG categories and assignments were obtained from the Joint Genome Institute's Integrate Microbial Genomes database.

C. Results and Discussion

Capture hybridization enrichment of *Vibrio* RNA. Capture hybridization baits were prepared by fragmenting *Vibrio* genomic DNA and transcribing the rRNA-depleted fragments into biotinylated RNA. We then made artificial mixtures of zebrafish and *Vibrio* RNA in order to test the efficacy of the hybridization capture method. Mixtures of 1%, 0.1% and 0.01% *Vibrio* RNA were prepared, and each mixture was divided into three parts and subjected to 0, 1, or 2 sequential pulldowns. Samples were pooled and sequenced on a single-end 150bp run on an Illumina HiSeq 4000. For each sample, the percent of the sequencing reads that aligned to the *Vibrio* reference genome and the fold-change compared to the pre-capture samples were calculated (Table 5-1).

Capture efficiency of defined mixtures of Vibrio and Zebrafish RNA

	Pre-Capture	One Capture		Two Capture	
Starting Fraction	Percent Aligned	Percent Aligned	Fold Change	Percent Aligned	Fold Change
1:100	0.80	29.8	38	69.7	88
1:1,000	0.07	6.0	89	38.3	568
1:10,000	0.01	0.9	110	10.0	1154

Table 5-1. Capture efficiency in defined ratios of *Vibrio* and zebrafish RNA. Percent Aligned is percentage of all reads passing filters that align uniquely to the genome. Fold change for each pulldown is by percent aligned and relative to pre-pulldown percent aligned for each sample.

In a 1:100 *Vibrio*:zebrafish mixture, *Vibrio* reads were enriched nearly 50-fold after two sequential captures. In the 1:1,000 mixture, reads were enriched more than 500-fold, making them the dominant type in the sequencing library. In the 1:10,000 mixture, *Vibrio* reads were enriched more than a thousand-fold, and comprised just under 10% of the final library.

We next tested the capture method using *in vivo* samples. Total RNA extracted from pools of larval zebrafish guts and a single gut was significantly enriched for *Vibrio* RNA, with 36-63% of reads aligning to the *Vibrio* genome after two rounds of capture hybridization (Table 5-2). This read alignment rate after enrichment is similar to what was seen with the artificial mixture of 0.1% *Vibrio*, consistent with the low *Vibrio*:host RNA ratio expected from *in vivo* samples.

Capture efficiency of Vibrio from larval zebrafish association

Host Genotype	Hours Post-Inoculation	Sample Type	Percent Aligned	Standard Deviation
Wild-Type	24	Pool	37	16
Wild-Type	72	Pool	62	6
Wild-Type	24	Single Gut	36	10

Table 5-2. Capture efficiency of host-associated Vibrio RNA after 2 sequential captures. Percent Aligned is percentage of all reads passing filters that align uniquely to the ZWU0020 genome. Sample type denotes whether the sample was prepared from a pool (5-10) larval zebrafish guts and associated Vibrio or a single larval zebrafish gut and associated Vibrio.

Hybridization capture does not bias expression data. Once the overall enrichment was determined, we then assessed if the capture hybridization biased the relative abundance of different transcripts. Variability in the capture efficiency for particular transcripts could lead to artifacts in differential expression analyses, and artificially increase or decrease the apparent expression of those genes within a particular condition. We compared RNA-Seq data both pre- and post-capture for the 1:100 *Vibrio*:zebrafish RNA artificial mixture. Raw counts of each gene pre- and post-capture correlate well (Figure 5-2 A). For most genes, there is a nearly 1:1 ratio of read counts in pre- and post-capture libraries, with an overall adjusted R^2 of 0.74. To further investigate potential bias, hierarchical clustering of these samples shows that the samples cluster by replicate, independent of the capture. This means that variation between biological replicates was greater than the variation pre-

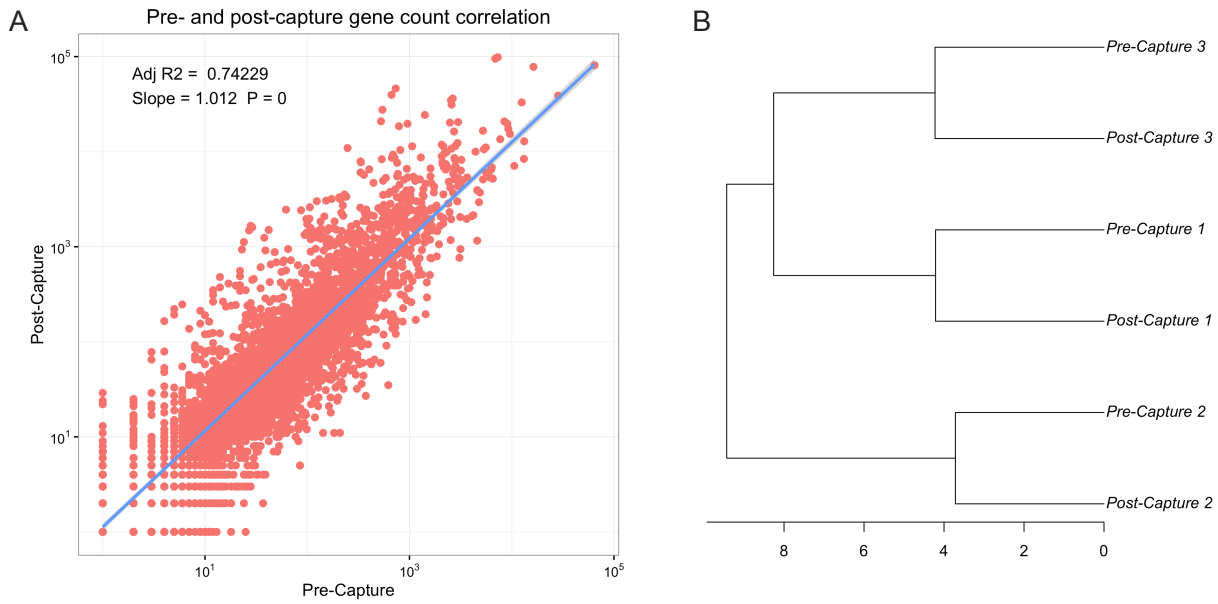

Figure 5-2. Hybridization capture expression data is unbiased. (A) Correlation of pre- and post- two capture raw gene counts. Raw gene counts pre- and post-capture of all genes in both conditions plotted on a log scale. (B) Hierarchical clustering of *in vitro* biological replicates. All samples group by biological replicate rather than pre-/post-capture.

and post-capture, demonstrating that our method does not impart significant bias in the data (Figure 5-2B).

***In vivo* and *in vitro* expression profiles differ significantly.** We first wanted to compare *in vivo* and *in vitro Vibrio* expression profiles. RNA-Seq libraries from 3 pools of larval zebrafish guts extracted 24 or 72 hours post *Vibrio* inoculation were subjected to two sequential capture hybridizations and sequenced. The resulting data was compared to *Vibrio* gene expression profiles from the mid-log *in vitro* cultures (post-capture) described above. Using DESeq2, we found vastly different *Vibrio* gene expression profiles between *in vitro* and *in vivo* RNA samples. Hierarchical clustering grouped samples into 3 distinct categories: *in vitro*, *in vivo* 24 hours post-inoculation, and *in vivo* 72 hours post-inoculation (Figure 5-3). Principal component analysis showed the major difference (98%) to be between *in vitro* and *in vivo* samples (data not shown).

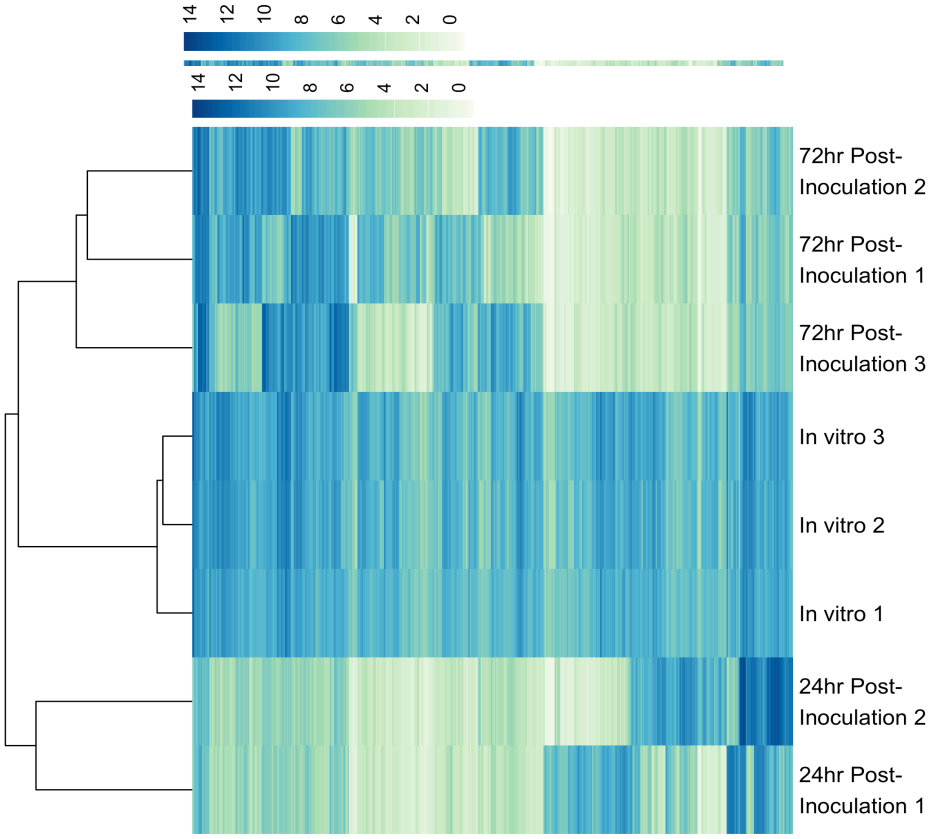

Figure 5-3. Expression changes *in vitro* and *in vivo*. Hierarchical clustering and expression of significantly differentially expressed genes *in vivo*. Blue indicates higher expression. Samples cluster into three distinct groups: *in vitro*, 24 hours post-inoculation, and 72 hours post-inoculation.

There were 1711 significantly differentially expressed genes (padj < 0.01), comprising ~50% of the predicted genes in the *Vibrio* genome, demonstrating that *Vibrio* exhibit a dramatic physiological shift in response to host colonization. The top 200 most differentially expressed genes were mapped to the genome (Figure 5-4).

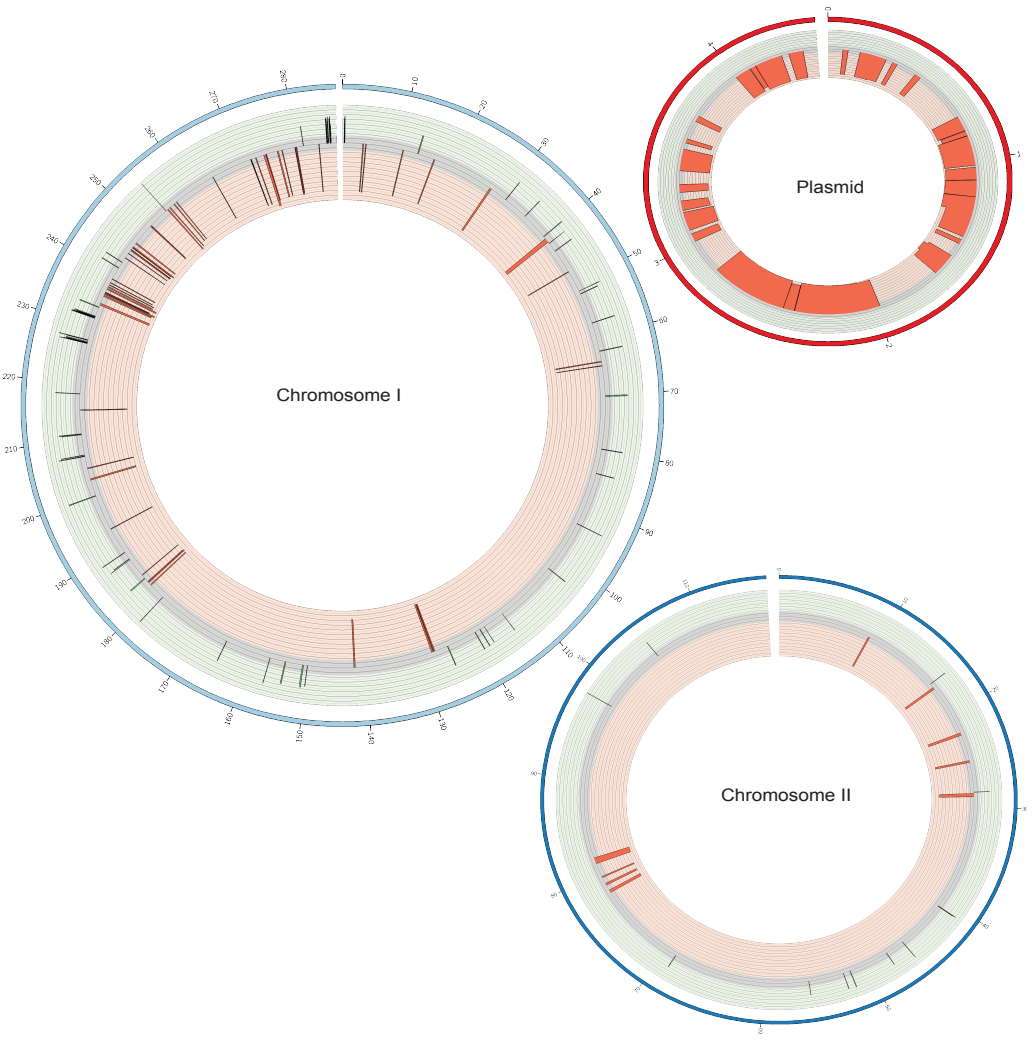

Figure 5-4. Distribution across ZWU0020 genome of 200 most highly differentially expressed genes between *in vitro* and *in vivo*. Log2 fold-change values are plotted along the length of their respective genes. Red bars represent negative log2 fold-changes and genes that are significantly downregulated *in vivo*. Green bars represent positive log2 fold-changes and genes that are significantly upregulated *in vivo*. Three putative plasmids from the genome assembly are not pictured.

In an effort to gain biological insight from the host-specific gene expression patterns, we focused on the 200 most differentially expressed genes. Many of the genes with the highest *in vivo* expression, relative to *in vitro*, are involved in functions anticipated to be important in the *in vivo* environment. Thirteen of them, have functions known to play a role in cellular stress. These include genes with roles in stress-induced cell envelope

integrity (ZWU0020_00738, ZWU0020_01348, ZWU0020_01847, and ZWU0020_02729), oxidative stress (ZWU0020_00268, ZWU0020_00532, and ZWU0020_01691), and a general stress response (ZWU0020_00354, ZWU0020_00850, ZWU0020_01072, ZWU0020_02829, ZWU0020_03453, and ZWU0020_03522). One of these genes, ZWU0020_01847, is a predicted homolog of the *yhcB* gene of *Vibrio fischeri*, which was shown to be important for squid colonization of this closely related bacterium (Brooks 2014). It is well known that host environments are depleted of bioavailable iron; consistent with this, we saw significant upregulation of several genes necessary for iron acquisition and utilization (ZWU0020_00475, ZWU0020_01677, ZWU0020_01679, and ZWU0020_02040). COG category analysis of the gene sets from the most highly differentially expressed genes and the whole genome shows enrichment

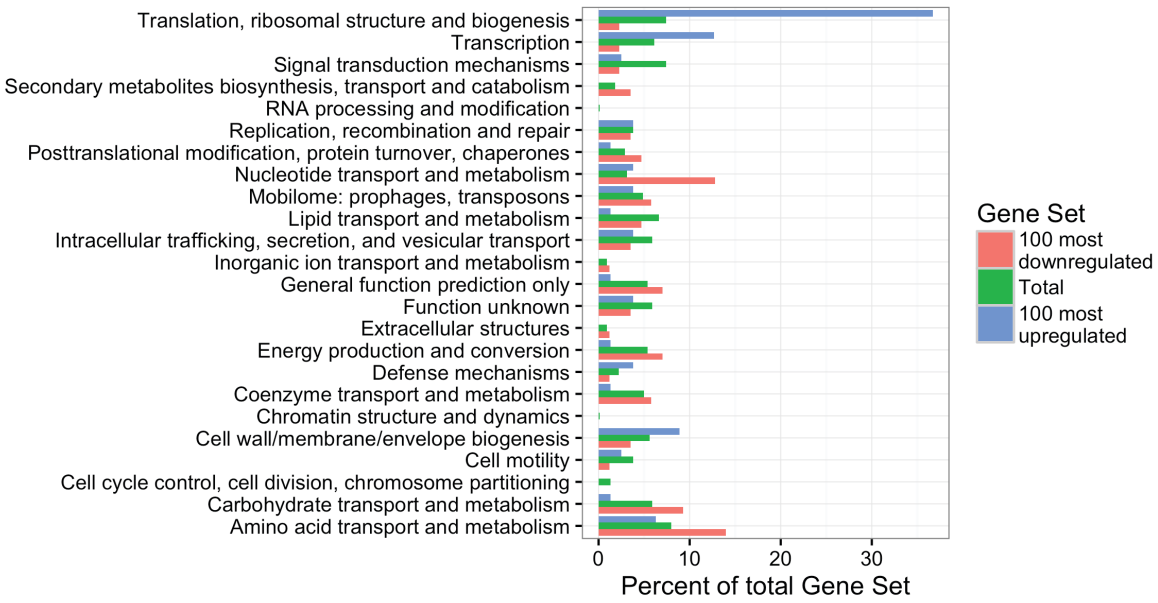

Figure 5-5. Distribution of COG category representation between *in vivo* and *in vitro* expression profiles. 100 most down regulated (pink) and upregulated genes (blue) *in vivo* compared to COG distribution of all genes in the Vibrio genome.

of specific COG categories in both the most unregulated and downregulated gene sets (Figure 5-5).

Genes involved in transcription and translation were particularly enriched in the most upregulated genes *in vivo*. Amino acid transport and metabolism genes were highly downregulated *in vivo*, which may be explained by the excess of peptones in the *in vitro* growth media. also mapped the , most of which mapped to Chromosome I. Based on the genome assembly, it is predicted to carry three plasmids, one of which is depicted in Figure 5-4. Interestingly, the majority of the genes on this plasmid are down regulated *in vivo*. The significance of this is unknown; however, gene expression levels of plasmid-encoded genes may be confounded by shifts in plasmid copy number. *Vibrio* genomes commonly harbor a large genetic island termed the superintegron, often comprised of hundreds of genes (Mazel 2006).

Our *Vibrio* isolate harbors a superintegron containing about 182 genes. By nature of how genes within superintegrons are integrated and regulated, these regions are generally transcriptionaly silent, as we see in our *Vibrio* RNA-seq data (Figure 5-6). This further verifies the validity of our *Vibrio* expression data and method.

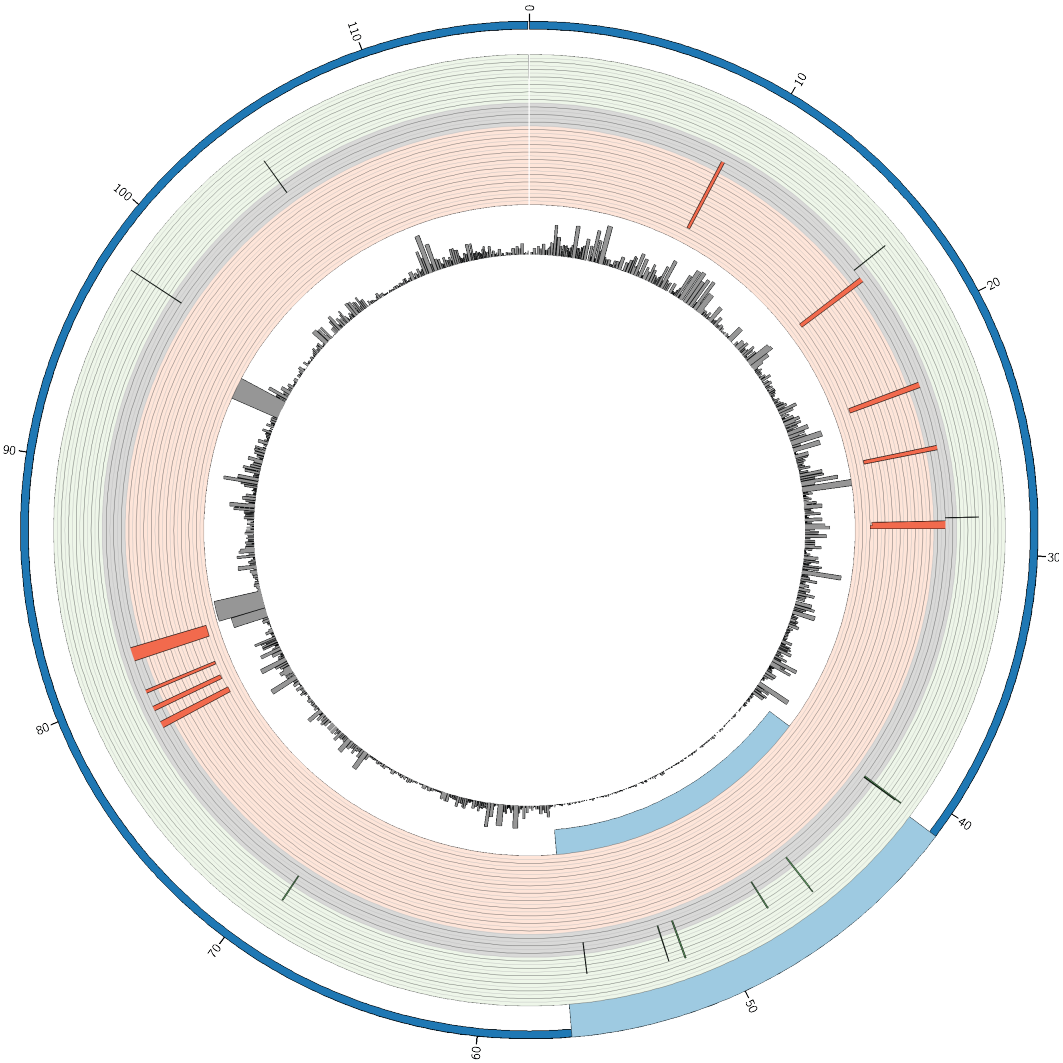

Figure 5-6. Depth of coverage on chromosome II of ZWU0020 with integron region. Integron region is highlighted in light blue and 200 most highly differentially expressed genes are shown as red and green bars. Depth of coverage for each gene is plotted on the innermost ring in grey. Red bars represent negative log2 fold-changes and genes that are significantly downregulated *in vivo*. Green bars represent positive log2 fold-changes and genes that are significantly upregulated *in vivo*.

Host-associated *Vibrio* exhibit temporal gene expression profile changes during host colonization. We captured Vibrio RNA from larval zebrafish guts collected 24 and 72 hours post-inoculation. Comparison of the 24hr and 72hr expression profiles revealed dramatic shifts in gene expression, including 311 significantly differentially expressed genes (padj < 0.05). The top 200 most differentially expressed genes were mapped to the

genome (Figure 5-7). Many more of the top most differentially expressed genes are located on chromosome II compared to the *in vitro* to *in vivo* comparison.

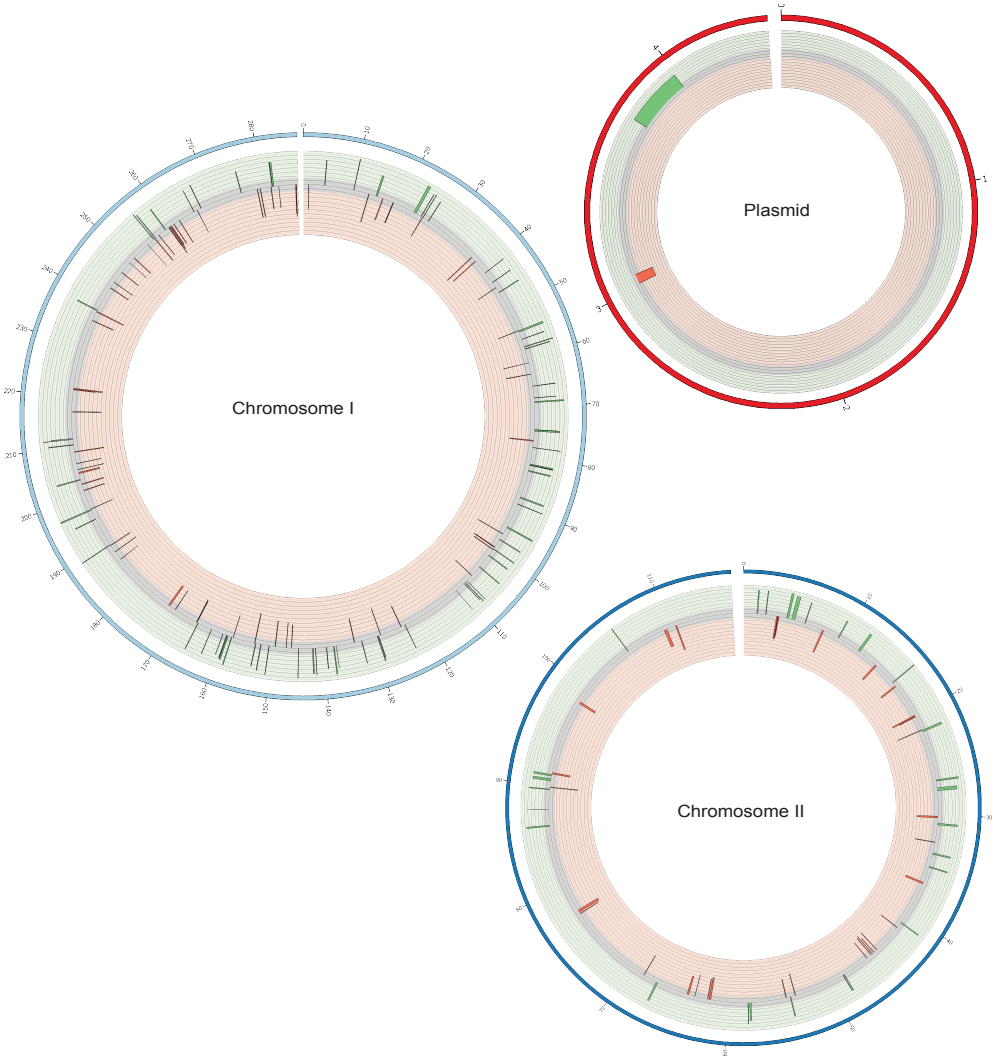

Figure 5-7. Distribution across ZWU0020 genome of 200 most highly differentially expressed genes *in vivo* between 24 and 72 hours post-inoculation. Log2 fold-change values are plotted along the length of their respective genes. Red bars represent negative log2 fold-changes and genes that are significantly downregulated 72 hours post-inoculation. Green bars represent positive log2 fold-changes and genes that are significantly upregulated 72 hours post-inoculation.

A previous study comparing the gene expression of *V. cholerae* in a rabbit ileal loop model to *in vitro* gene expression saw a shift to more small chromosome (II) genes expressed *in vivo* (Xu 2003). Although we did not see this in our *in vitro* to *in vivo* comparison, perhaps in these conditions there is a delay in this shift such that it is not evident until longer *in vivo* colonization. Furthermore, in these data there were few significantly differentially expressed genes on the plasmid for which we saw a dramatic down regulation when comparing *in vivo* to *in vitro* expression data (Figure 5-4). Signatures of specific physiological changes ascribed by the top differentially expressed genes are unclear in these data. Included on the top upregulated gene list, however, are genes that are expected to be induced *in vivo*, including collagenase (ZWU0020_00861), hemolysin (ZWU0020_02874), and the transcriptional activator VirB (ZWU0020_02676). These temporal changes in gene expression could be the result of the length of time of colonization independent of the state of host development; however, it cannot be discounted that these changes may also be due to changes in the host environment resulting from developmental changes in the fish larvae, regardless of the duration of colonization. Indeed, the larval fish is experiencing rapid development during these early days of life. Studies designed to disentangle the effects of these two factors is warranted.

Differentially expressed genes agree with previous studies.

Stephens *et. al* (2015) used a transposon insertion knockout library of *Vibrio ZWU0020* to investigate which genes are important for colonization of the zebrafish gut. They were

able to assay 1930 genes (~50% of predicted genes) and found 278 high-confidence gene candidates that potentially play a role in host colonization. In order to facilitate comparison of the results of their study to ours, we ran their list of high-confidence genes, and our list of significantly differentially expressed genes from this study, through GAGE analysis. This KEGG-based analysis showed two KEGG orthology pathways to be enriched in the Stephens *et al*. study for Vibrio mono-association, including flagellar assembly, and two-component signal-transduction systems. In our study, two-component signaling systems were also significantly enriched in expression profiles for *in vivo Vibrio*, demonstrating that our results at least partially recapitulate the findings of the previous stud. Indeed, comparison of these two different types of genetic screens is difficult due to the inherent differences how they probe genetic functional importance. RNA-Seq experiments allow for assessing the role of a larger portion of the genes in an organism than transposon mutagenesis studies. It is also possible to assay the significance of genes that are required for growth in both the *in vitro* and *in vivo* conditions being compared (i.e. essential genes), as these would not be present in the pool of transposon library mutants. Our study assayed all annotated genes and identified 1,711 significantly differentially expressed genes, compared to 278 candidate genes in the Stephens *et al* study. Previous work comparing transposon-insertion data to transcriptomic data found a poor correlation (~3% of all genes) between the two data sets (Powell, 2016). We found better overlap- about 7% of all genes were found to be important in host colonization in both transposon sequencing and by RNA-Seq.

D. Conculsion

Hybridization capture enriches for bacterial RNA-Seq libraries. The RNA capture method presented here is highly efficient at enriching for bacterial RNA in host-associated samples where the bacterial RNA is an extremely minor component of the total RNA. This method is widely applicable for studying gene expression in other host-microbe systems where microbial constituents are difficult to isolate from host tissues. Moreover, this method could prove a powerful tool in other applications where the target species is too closely associated with another more dominant species for separation, as with intracellular parasites. Because total RNA-Seq libraries are made prior to capturing the bacterial portion, it is possible to sequence the pre-capture libraries to obtain gene expression data from the host tissue. Host and microbe gene expression data from the same sample would give unique insights into host-microbe interactions. We were able to generate reliable gene expression data from mono-associated larval zebrafish, but the method could be expanded to isolate more complex communities. Bait libraries can be generated from any culturable microbe and then combined to enrich for all microbial RNA from samples isolated from a host inoculated with a defined community. Combining this method with *de novo* transcriptome assembly could allow discovery of novel transcripts in less well-characterized microbes.

Table 5-3. Strategies for enriching rare RNA

Strategy	Advantages	Disadvantages	Citations
Mechanical Separation (Size-based Filtering, Selective Lysis)	Removes host RNA before bacterial lysis	Too much sample handling Unsuitable for intracellular targets	Lim, *et al* 2012
Host RNA Depletion (CpG removal, Poly-A removal)	Easy, commercially available kits	Inefficient for rare targets	Kumar, *et al* 2016
Hybridization capture-based methods (PCR-generated probes,	targeted probe to specific region	Too little information Too laborious to reduce representation bias	Faucher, *et al* 2005 An, *et al* 2012
cDNA generated probes) Micro-fluidic capture of cDNA	Specific to entire genome of interest	Need to have culturable bacteria Complicated isolation Only one sample at a time	Bent, *et al* 2013
Our Method	Specific to entire genome of interest	Need to have culturable bacteria	

Previous efforts to sequence rare RNA are summarized in Table 5-6. They fall into several broad categories: mechanical separation, host RNA depletion, microfluidic capture, and hybridization-based capture. Mechanical separation involves size-based filtration or differential lysis of prokaryotic and eukaryotic cells, but sample handling time is significantly longer than bacterial mRNA turnover rates and transcriptome profiles are influenced by the sample preparation. Host RNA depletion methods are convenient as there are several commercially available kits for CpG island and Poly-A depletion, but are less effective when target RNA is very rare. Kumar et. al. describe a method for poly-A depletion of eukaryotic transcripts to enrich for Wolbachia transcripts from Wolbachia-infected Drosophila. Poly-A depletion and bacterial rRNA depletion increased bacterial mRNA 3-fold, but final libraries contained 1.0% bacterial mRNA, which is not sufficient for transcriptomic analysis. Microfluidic capture of cDNA with probes generated from genomic DNA was much more successful. Bent et. al. were able to enrich several hundred fold for RNA from the intracellular parasite *Francisella Tularensis*. However, their method involved specialized microfluidic equipment and is not high throughput, as individual captures must be performed for each sample. Previous hybridization-based capture methods have used PCR amplicons and cDNA as templates

for bait library, which does not allow for interrogation of the entire transcriptome and requires laborious steps to normalize counts across different genes. By using gDNA as the bait library template, we were able to increase the ratio of microbial:host RNA by more than one thousand-fold and obtain usable transcriptomic data for differential expression analysis. By capturing pooled libraries with streptavidin beads, we were able to create a relatively easy, scalable process for enriching bacterial trancsripts. We were able to demonstrate that the capture method does not bias the transcriptome data, since variation between biological replicate cultures was significantly greater than that between pre- and post-capture replicates. Furthermore, this method enables preparing total RNA from the samples immediately upon dissection, so that the potential for changes in gene expression profiles during an otherwise needed bacterial isolation step is minimized.

CHAPTER VI

CONCLUSION

Highly heterogeneous populations are by nature particularly difficult to characterize. Next-generation sequencing (NGS), is powerful tool to investigate homogeneous populations, but is less useful in investigating complex populations. Rare sequence variant identification is confounded by the error rate of sequencing instruments. To address this issue, I have co-developed a method described in Chapter II to improve the error rate of NGS. This method was used to characterize mutations arising during tumorigenesis and the spatial distribution of mutations within a solid tumor in Chapter III. With our new, sensitive detection of rare variants we were able to see distinct patterns of variation in different regions of the mitochondrial genome in tumor and non-tumor cells from the same subject, potentially because of different selective pressures on each region. We were also able to detect spatial organization in different regions of a solid tumor. Because most of the variants in the tumor are unique to a specific section, it appears that the overall mutation rate within the tumor is high. PELE-Seq was also used to investigate mutation accumulation in fanconi anemia in Chapter IV. We determined that SNPs and small (<30bp) indels are more frequent in WT than in fanconi mutant tissues, potentially due to impaired error-prone polymerase recruitment in the mutant fish.

Rare members in a complex biological community present a different challenge, such as in characterizing host-microbe interactions. I co-developed a method for enriching for bacterial transcripts from host-associated bacteria described in Chapter V. This method increased the ratio of microbial:host RNA by more than one thousand-

fold and produced usable transcriptomic data for differential expression analysis. By capturing pooled libraries with streptavidin beads, we were able to create a relatively easy, scalable process for enriching bacterial trancsripts. The capture method does not bias the transcriptome data and it enables preparing total RNA from the samples immediately upon dissection, so that the potential for changes in gene expression is minimized.

Together, these methods allow for more thorough investigation of complex communities. They create possibilities for investigating complex communities that would previously have been obscured by errors and noise.

References Cited

Kaiser J. The Downside of Diversity. Science. 2013;339(6127):1543-1545.

Bhatia S, Frangioni, J, Hoffman R, Iafrate AJ, Polyak K. The challenges posed by cancer heterogeneity. Nature Biotechnology. 2012;30:604–610.

Modi S, Lee H, Spina C, and Collins J. Antibiotic treatment expands the resistance reservoir and ecological network of the phage metagenome. Nature. 2013;499:219-222.

Hohenlohe P, Bassham S, Etter P, Stiffler N, Johnson EA, Cresko W. Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags. PLOS Genetics. 2010;6(2):e1000862.

Nielsen R, Paul JS, Albrechtsen A, Song YS. Genotype and SNP Calling from Next-Generation Sequencing Data." Nature reviews. Genetics. 2011; 12.6: 443–451.

Marçais G, Yorke JA, Zimin A. QuorUM: an error corrector for Illumina reads. PLoS One. 2015;10(6): e0130821.

Schloissnig S, Arumugam M, Sunagawa S, Mitreva M, Tap J, Zhu A, et al. Genomic variation landscape of the human gut microbiome. Nature. 2013;493,45–50.

Kircher M, Kelso J. High-throughput DNA sequencing - concepts and limitations. Bioessays. 2010;32:524-536.

Goto H, Dickins B, Afgan E, Paul IM, Taylor J, Makova MD, et al. Dynamics of mitochondrial heteroplasmy in three families investigated via a repeatable re-sequencing study. Genome Bio. 2011;12:R59.

Osvaldo Zagordi, Rolf Klein, Martin Däumer, and Niko Beerenwinkel. Error correction of next-generation sequencing data and reliable estimation of HIV quasispecies. Nucleic Acids Res. 2010; gkq655v1-gkq655.

Chen-Harris H, Borucki M, Torres C, Slezak T, Allen J. Ultra-deep mutant spectrum profiling: improving sequencing accuracy using overlapping read pairs. BMC Genomics. 2013;14:96.

Maura Costello, Trevor J. Pugh, Timothy J. Fennell, Chip Stewart, Lee Lichtenstein, James C. Meldrim, et al. Discovery and characterization of artifactual mutations in deep coverage targeted capture sequencing data due to oxidative DNA damage during sample preparation. Nucleic Acids Res. 2013 Apr; 41(6): e67.

Jeong H, Barbe V, Lee C, Vallenet D, Yu D, Choi S, et al. Genome sequences of Escherichia coli B strains REL606 and BL21(DE3). J Mol Biol. 2009;4:644-52.

Hayashi K, Morooka N, Yamamoto Y, Fujita K, Isono K, Choi S, et al. Highly accurate genome sequences of Escherichia coli K-12 strains MG1655 and W3110. Mol Syst Biol. 2006;2:2006.0007.

Sikkink K, Reynolds R, Ituarte C, Cresko W, Phillips P. Rapid evolution of phenotypic plasticity and shifting thresholds of genetic assimilation in the nematode Caenorhabditis remanei. G3: Genes, Genomes and Genetics. 2014;4:1103-1112.

Wilm A, Aw P, Bertrand D, Yeo G, Ong S, Wong C, Khor, C, et al. LoFreq: a sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets. Nucleic Acids Res. 2012;22:11189-111201.

Baird N, Etter P, Atwood T, Currey M, Shiver A, Lewis Z, et al. Rapid SNP discovery and genetic mapping using sequenced RAD markers.” PLoS One. 2008;3(10):e3376.

Alexandrov L, Nik-Zainal S, Wedge DC, Aparicio SA, Behjati S, Biankin AV,et al. "Signatures of mutational processes in human cancer" Nature. 2013;500(7463):415-421.

Pfeifer GP. Mutagenesis at methylated CpG sequences. Curr Top Microbiol Immunol. 2006;301:259-81.

Brodin J, Mild M, Hedskog C, Sherwood E, Leitner L, Andersson B. PCR-Induced Transitions Are the Major Source of Error in Cleaned Ultra-Deep Pyrosequencing Data. PLoS ONE 8(7): e70388.

Christoforides A, Carpten JD, Weiss GJ, Demeure MJ, Von Hoff DD, Craig DW. Identification of somatic mutations in cancer through Bayesian-based analysis of sequenced genome pairs. BMC Genomics, 2013;14(1):1.

Peterson BK, Weber JN, Kay EH, Fisher HS, Hoekstra HE. Double Digest RADseq: An Inexpensive Method for De Novo SNP Discovery and Genotyping in Model and Non-Model Species. PLoS ONE. 7(5): e37135

Pan L, Shah AN, Phelps IG, Doherty D, Johnson EA and Moens CB. Rapid identification and recovery of ENU-induced mutations with next-generation sequencing and Paired-End Low-Error analysis. BMC Genomics. 2015;16:8.3

Gibson G. Rare and common variants: twenty arguments. Nature Reviews Genetics. 2012;13, 135-145.

De La Vega FM, Bustamante CD, Leal SM. Genome-wide association mapping and rare alleles: from population genomics to personalized medicine. Pac Symp Biocomput. 2011:74-5.

King CD, Rios GR, Green MD, Tephly TR. UDP-Glucuronosyltransferases. Current Drug Metabolism. 2000;19:143-161.

Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, et al. Circos: an information aesthetic for comparative genomics. Genome Research. 2009;19:1639-1645.

Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, et al. Integrative Genomics Viewer. Nature Biotechnology. 2011;29: 24–26.

Thorvaldsdóttir H, Robinson JT, Mesirov JP. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Briefings in Bioinformatics. 2013;14:178-192.

Phanstiel DH. Sushi: Tools for visualizing genomics data. R package version 1.8.0., 2015.

Wickham H. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2009.

Mardis ER.DNA sequencing technologies: 2006–2016.Nature Protocols. 2017;12: 213–218.

Wall JD, Tang LF, Zerbe B, Kvale MN, Kwok PY, Schaefer C, Risch N. Estimating genotype error rates from high-coverage next-generation sequence data. 2014;11:1734-9.

Payne BAI, Wilson IJ, Yu-Wai-Man P, Coxhead J, Deehan D, Horvath R, Taylor RW, Samuels DC, Santibanez M, Chinnery PF. Universal heteroplasmy of human mitochondrial DNA. Hum Mol Genet. 2013;2:384-390.

Chatterjee A, Mambo E, Sidransky D. Mitochondrial DNA nutations in human cancer. Oncogene 2006;24:4663-74.

Araten DJ, Golde DW, Zhang RH, Thaler HT, Gargiulo L, Notaro R, Luzzatto L. A quantitative measurement of the human somatic mutation rate. Cancer Res. 2005;22:10635.

Neish AS. Microbes in gastrointenstinal health and disease. Gastroenterology. 2009;136:65-80.

Wiles TJ, Jemielita M , Baker RP, Schlomann BH, Logan SL, Ganz J, Melancon E, Eisen JS, Guillemin K, Parthasarathy R. Host Gut Motility Promotes Competitive Exclusion within a Model Intestinal Microbiota. PLOS Biology. July 26, 2016.

Edgar, R.C. UPARSE: Highly accurate OTU sequences from microbial amplicon reads, Nature Methods. 2013

Li H. and Durbin R. Fast and accurate short read alignment with Burrows-Wheeler Transform. Bioinformatics. 2009;25:1754-60.

Wilm et al. LoFreq: A sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets. Nucleic Acids Res. 2012; 40(22):11189-201.

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker R, Lunter G, Marth G, Sherry ST, McVean G, Durbin R and 1000 Genomes Project Analysis Group. The Variant Call Format and VCFtools. Bioinformatics, 2011.

McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GR, Thormann A, Flicek P, Cunningham F. The Ensembl Variant Effect Predictor. Genome Biology 2016;17:122.

Carpenter ML, Buenrostro JD, Valdiosera C, Schroeder H, Allentoft ME, Sikora M, et al. Pulling out the 1%: Whole-Genome Capture for the Targeted Enrichment of Ancient DNA Sequencing Libraries. Am. J. Hum. Genet. 2013;93:852–64.

Stephens WZ, Wiles TJ, Martinez ES, Jemielita M, Burns AR, Parthasarathy R, Bohannan BJ, Guillemin K. Identification of Population Bottlenecks and Colonization Factors during Assembly of Bacterial Communities within the Zebrafish Intestine. MBio. 2015;6:1163-15.

Rolig AS, Parthasarathy R, Burns AR, Bohannan BJM, Guillemin K. Individual members of the microbiota disproportionately modulate host immune responses. Cell Host and Microbe. 2015;18:613-620.

Hill JH, Franzosa EA, Huttenhower C, Guillemin K. A conserved bacterial protein induces pancreatic beta cell expansion during zebrafish development. eLife 2016;5:e20145.

Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26:841-42.

Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology. 2014;15:550.

Luo W, Friedman MS, Shedden K, Hankenson KD, Woolf PJ. GAGE: generally applicable gene set enrichment for pathway analysis. BMC Bioinformatics. 2009;10:161.

Luo W, Brouwer C. Pathview: an R/Bioconductor package for pathway-based data integration and visualization. Bioinformatics. 2013;29:1830-31.

Brooks JF, Gyllborg MC, Cronin DC, Quillin SJ, Mallama Ca, Foxall R, Whistler C, Goodman AL, Mandel MJ. Global discovery of colonization determinants in the squid symbiont *Vibrio fischeri*. PNAS. 2014;111:17284-17289.

Mazel D. Integrons: agents of bacterial evolution. Nat Rev Microbiol. 2006;4:608-20.

Xu Q, Dziejman, Mekalanos JJ. Determination of the transcriptome of *Vibrio cholerae* during intraintestinal growth and midexponential phase *in vitro.* PNAS. 2013;100:1286-1291.

Powell JE, Leonard SP, Kwong WK, Engel P, Moran NA. Genome-wide screen identifies host colonization determinants in bacterial gut symbiont. PNAS. 2016;113:13887-13892.

Kumar N, Lin M, Zhao X, Ott S, Santana-Cruz I, Daugherty S et al. Efficient Enrichment of Bacterial mRNA from Host-Bacteria Total RNA Samples. Sci Rep. 2016;6:34850.

Faucher SP, Curtiss R, Daigle F. Selective capture of Salmonella enterica serovar typhi genes expressed in macrophages that are absent from the Salmonella enterica serovar Typhimurium genome. Infect Immun. 2005;73:5217-21.

An R, Grewal PS. Selective Capture of Transcribed Sequences: A Promising Approach for Investigating Bacterium-Insect Interactions. Insect. 2012;3:295-306.

Bent ZW, Brazel DM, Tran-Gyamfi MB, Hamblin RY, VanderNoot VA, Branda SS. Use of a capture-based pathogen transcript enrichment strategy for RNA-Seq analysis of the Francisella Tularensis LVS transcriptome during infection of murine macrophages. PLoS One. 2013;8(10):e77834