OVERCOMING THE CURRENT LIMITATIONS OF NEXT-GENERATION SEQUENCING WITH NEW METHODS FOR LOCAL ASSEMBLY OF GENOMES AND HIGH-SPECIFICITY RARE MUTATION DETECTION

by

JESSICA L. PRESTON

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

December 2015

DISSERTATION APPROVAL PAGE

Student: Jessica L. Preston

Title: Overcoming the Current Limitations of Next-Generation Sequencing with New Methods for Local Assembly of Genomes and High-Specificity Rare Mutation Detection

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:

Patrick Phillips	Chairperson
Eric Johnson	Advisor
Bruce Bowerman	Core Member
John Postlethwait	Core Member
Diane Hawley	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 December 2015

DISSERTATION ABSTRACT

Jessica L. Preston

Doctor of Philosophy

Department of Biology

December 2015

Title: Overcoming the Current Limitations of Next-Generation Sequencing with New Methods for Local Assembly of Genomes and High-Specificity Rare Mutation Detection

The relatively low cost of Next-Generation Sequencing (NGS) has enabled researchers to generate large amounts of sequencing data in order to identify disease-causing mutations and to assemble simple genomes. However, NGS has inherent limitations due to the short DNA read lengths and high error rate associated with the technique. The short read lengths of NGS prevent the assembly of genomes with long stretches of repetitive DNA, and the high error rate prevents the accurate detection of rare mutations in heterogeneous populations such as tumors and microbiomes.

I have co-developed new NGS methods to overcome these challenges. In order to increase the effective read length of NGS reads, local *de novo* assembly of short reads into long contigs can be achieved through the use of Paired-End Restriction-site Associated DNA Sequencing (RAD-PE-Seq). With the RAD-PE method, I sequenced a stickleback fosmid and generated contigs with an N_{50} length of 480 nucleotides. In order to eliminate false-positive mutations caused by the high error rate of NGS, the Paired-End Low Error Sequencing (PELE-Seq) method was developed, which uses numerous quality control measures during the sequencing library preparation and data analysis steps

in order to effectively eliminate sequencing errors. Control testing of the PELE-Seq demonstrates that the method completely eliminates false-positive mutations at sequencing read depths below 20,000X coverage, compared to a ~20% false-positive rate obtained with previous methods. The high accuracy of the PELE-Seq method allows for the detection of ultra-rare mutations in a genome, which was previously impossible with NGS.

This dissertation includes previously published and unpublished co-authored material.

CURRICULUM VITAE

NAME OF AUTHOR: Jessica L. Preston

GRADUATE AND UNDERGRADUATE SCHOOLS ATTENDED:

University of Oregon, Eugene, OR
Humboldt State University, Arcata, CA

DEGREES AWARDED:

Doctor of Philosophy, Biology, 2015, University of Oregon
Bachelor of Science, Chemistry, 2008, Humboldt State University

AREAS OF SPECIAL INTEREST:

Functional Genomics
DNA Sequencing
Tumor Evolution

PROFESSIONAL EXPERIENCE:

Laboratory Technician, Hui Zong Lab, University of Oregon, 2008-2009

Undergraduate Researcher, Athar Chishti Lab, University of Illinois at Chicago, 2007

GRANTS, AWARDS, AND HONORS:

Best Poster, EVO-WIBO Conference, "Sequencing Rare Alleles in Ancestral and Laboratory-Adapted Strains of *C. remanei* with Paired-End Low-Error Sequencing (PELE-Seq)."Port Townsend, WA, 2014

UO Cancer Federation Scholarship, 2013

UO Women in Graduate Sciences Parenting Scholarship, 2012

PUBLICATIONS:

Paul D. Etter, Jessica L. Preston, Susan Bassham, William A. Cresko & Eric A. Johnson. Local de novo Assembly of RAD Paired-end Contigs Using Short Sequencing Reads. PLoS ONE. 2011, Apr13; 6(4):e18561.PMID:21541009

ACKNOWLEDGMENTS

I wish to express sincere appreciation to Professor Eric A. Johnson for his assistance in the preparation of this manuscript. In addition, special thanks are due to Dr. Paul Etter and Dr. Doug Turnbull, whose familiarity with current Next-Generation Sequencing techniques was helpful during the development phase of this undertaking. Thank you to Professor Hui Zong for his invaluable mentorship and to Dr. Rui Galvao and Professor Chong Liu for providing MADM mouse tumor samples. Thank you to Professor Patrick Phillips and Dr. Kristin Sikkink for providing lab-adapted strains of *C. remanei*. Thank you to Professor Bill Cresko and Professor Julian Catchen for their support with data analysis. Thank you also to Dr. Paul Spellman's lab for providing exome sequencing data for the PELE-Seq blood biopsy project. The investigation was supported in part by a Public Health Service Predoctoral Fellowship, Number 5 T32 GM 7413-35, from the National Institutes of Health.

Dedicated to my grandfather Donald Olsen.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Table Page

CHAPTER I

INTRODUCTION

The cost of genome sequencing has decreased dramatically in the last decade, and it is now possible to sequence a human genome for a few thousand dollars in a matter of days. Next-Generation DNA Sequencing (NGS) is a massively parallel, high-throughput process that generates sequence information for hundreds of millions of DNA molecules simultaneously. NGS has revolutionized molecular genetics, leading to the creation of the fields of transcriptomics and metagenomics, as well as others. The increased affordability of DNA and RNA sequencing has led to the generation of massive amounts of sequence data and new avenues of research. However, NGS is still a relatively new technique with some major limitations, namely, the short length of the sequencing reads (less than 500 nucleotides) and the high error rates of the sequencer (~1%) [1,2,3]. The short read lengths of NGS platforms make it difficult to assemble genomes that contain highly repetitive regions, or to properly differentiate between alternative RNA isoforms[4,5]. The high error rates of the sequencing platform software and the polymerase enzymes used have led to an extremely high false positive rate, which makes it extremely challenging to accurately detect rare mutations present in a heterogeneous population, such as tumors or mitochondria [6,7]. The problems resulting from the short read lengths and high error rates associated with NGS pose challenges for researchers studying diverse topics ranging from evolutionary to clinical biology. This dissertation describes work done to address these problems through the development of new methods and techniques to overcome the inherent limitations of NGS. The first section describes a method

designed to overcome the problems associated with the short sequencing read length through the local assembly of short reads using restriction enzyme sites. The second section describes a method to reduce the error rate of NGS sequencing data using overlapping paired-end (PE) reads and a barcoding system.

Local Assembly of Short Reads Using Restriction Sites

One major shortcoming of the current high-throughput DNA sequencing methods is the short length of the sequence reads generated by the sequencing platforms [4,5]. For many applications, the short read length is not problematic and sequencing at a high read depth can generate enough information to assemble the short reads into contiguous sequences (contigs). Contigs are generated with programs that find where short sequencing reads overlap each other and then use that information to assemble them into one sequence. However, short sequencing read lengths are problematic in a variety of circumstances. Short read lengths interfere with genome assembly when the genome contains long stretches of repetitive regions, which is the case with most large genomes. If the stretches of repetitive sequences present in the genome are larger than the sequencing reads it is impossible to generate contigs by simply overlapping the sequence reads, as there is no way to know where the repetitive sequences begin and end. Another limitation resulting from short sequencing reads is the inability to infer information about the number and type of mRNA isoforms that have resulted from alternative splicing. Because each sequencing read is shorter than the length of a full mRNA transcript, it is impossible to piece together the various isoforms that were originally present in the

sample. Similarly, short sequencing read lengths also make it challenging to distinguish which homologous chromosome a sequencing read belongs to, hence it is difficult to determine haplotype information from a sample [3].

I have co-developed a new method to assemble short sequencing reads based on their position relative to restriction enzyme sites. Because restriction enzyme sites are found at specific sites throughout a genome, the short reads generated by sequencing each site can be grouped together and assembled locally into contigs. In order to generate longer contigs, it is useful to sequence as much of the genome space next to a restriction site as possible. The method, called RAD Paired-End Sequencing (RAD-PE Seq) is described in Chapter II, which is reproduced with permission from Etter PD, Preston JL, Bassham S, Cresko WA, Johnson EA, 2011, 6(4): e18561. Copyright 2011, *PLoS ONE.*

Identifying and Filtering Next-Generation Sequencing Errors

It is currently very challenging to detect rare mutations in genetically heterogeneous populations such as tumors and microbiomes with NGS, due to the relatively high error rates of NGS sequencing platforms and polymerase enzymes [6,7]. When sequencing a sample at 100x depth of coverage, the 1% error rate of current Illumina sequencers can lead to the generation of one error at every position in the genome. These sequencing and PCR errors are difficult to distinguish from true rare genetic variants using current methods. I have co-developed a new variant-calling method that drastically reduces the incidence of sequencing errors, called Paired-End Low Error Sequencing (PELE-Seq). The PELE-Seq method is described in Chapter III, which has

been submitted for publication 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)” (*submitted to BMC Genomics*).

CHAPTER II

LOCAL GENOME ASSEMBLY OF SHORT SEQUENCING READS USING RESTRICTION SITES

This work was published as Etter PD, Preston JL, Bassham S, Cresko WA, Johnson EA. Local *De Novo* Assembly of RAD Paired-End Contigs Using Short Sequencing Reads. 2011, 6(4): e18561. Eric Johnson and Paul Etter developed the RAD-PE method. I generated the data (partial-digest RAD-PE in stickleback fosmids). Eric Johnson performed data analysis. Eric Johnson and Bill Cresko were the principal investigators for this work.

Introduction

Despite the power of massively parallel sequencing platforms, *de novo* assembly of genomes with the short reads produced remains difficult. We demonstrate that short reads can be locally assembled into larger contigs using paired-end sequencing of restriction-site associated DNA (RAD-PE) fragments. We use this RAD-PE contig approach to sequence *E. coli* and stickleback genomic DNA with overlapping contigs of several hundred nucleotides. RAD-PE contigs mitigate the problem of short reads by creating much longer high-quality contigs appropriate for SNP discovery or the de novo assembly of genomes.

The decreased cost and throughput increases offered by next-generation sequencing platforms create the ability to produce high coverage of a genome in a short

time. However, it remains difficult to move from many millions of short reads to a high-quality assembled genome, as the short sequence read lengths and the high error rates create computational difficulties. Several algorithms have been developed to more efficiently work with short read datasets [1,2,3], but these approaches require costly computing resources to compare each sequence read against all others [1,2,3,4].

One difficulty in assembling a genome from short reads is bridging repetitive sequences. These sequences may exist in thousands to millions of locations in a genome, and are nearly indistinguishable in the context of a short sequence read. Without a way to place each repetitive sequence in its proper genomic location, it is difficult to move beyond producing a genome sequence made of many shorter contigs. Traditional solutions to this problem have included physically breaking the genome into smaller fragments, then cloning and sequencing each fragment independently, thereby ensuring that each repetitive sequence can be localized to a small region of the genome. The complexity reduction created by physically isolating a shorter genomic fragment is laborious, but remains one of the few true solutions to the challenges of assembling a complex genome.

RAD tags are based on a different sort of complexity reduction step that samples the DNA flanking each instance of a particular restriction site in the genome [5,6,7,8]. RAD tags were developed to speed discovery of SNPs and have been particularly attractive in systems lacking a reference genome. However, moving from SNPs identified by sequencing RAD tags to a high-throughput genotyping platform is difficult without a reference genome, as these platforms typically require more than 60 nucleotides of flanking genomic DNA on both sides of the SNP of interest.

A distinctive feature of RAD tags is the asymmetric nature of the DNA fragments. Each RAD tag has one end defined by the restriction enzyme recognition site, and the other end defined by random shearing. Next-generation sequencers now have the capability to carry out paired-end reads, in which the two ends of a DNA fragment are sequenced and the two end sequences are known to belong to the same fragment. Paired-end sequencing enables RAD fragments to be used for local de novo assembly. A typical RAD library may contain 10,000 to 100,000 RAD sequences. The sheared-end sequences that share a common RAD-site sequence are all derived from the same small region near the RAD site. This small set of sheared-end sequences can be assembled into a larger contig. Instead of a single, computationally intense assembly using the many sequence reads from the entire genome, RAD paired-end contig assembly is performed using only a small portion of the data at a time. Because the sequence reads come from a small region, the difficulties of finding significant sequence overlap and dealing with sequence errors become simpler. To demonstrate the power of this approach, we have created created RAD-PE contigs after a partial digest with a restriction enzyme that cuts at high frequency to generate overlapping contigs in stickleback.

RAD Paired-End Contig Library Generation

The DNA fragments created by RAD tag library preparation have a restriction site at one end and are randomly sheared at the other. This arrangement, when combined with Illumina paired-end sequencing, results in each instance of a restriction site sequence being sampled many times by the first reads and the genomic DNA sequence in the

nearby region being randomly sampled at a lower coverage by the second reads. We hypothesized that the explicit linking of second reads that sample a genomic region with a common first read RAD sequence would allow the second reads to be assembled on a local basis, one RAD site at a time (see Figure 2.1).

We tested this approach by modifying the sequenced RAD tag protocol [6] in order to create paired-end compatible libraries. We altered two key aspects of the RAD protocol. First, a wider size range of fragments (300-800 bp) was isolated after shearing. The size of contigs assembled from the paired-end reads is dependent on the size range of fragments selected during library construction. Second, a longer, divergent P2 adapter that contains the reverse sequencing primer sequence was ligated to the variable end of the RAD tags before amplification, allowing the randomly sheared end of the RAD fragments to be sequenced by the second read. In order to make RAD-PE contigs useful for de novo assembly of whole genomes, and to achieve high coverage of a whole genome, libraries were created by partially digesting with a high-frequency restriction enzyme, which produced overlapping DNA fragments several kb long that were suitable for shearing (Figure 2.2A). As a result, RAD cut sites are typically only a few hundred base pairs apart, but the sheared ends sample 500 bp regions to the left and right of each RAD site.

Partial-Digest RAD Paired-End Libraries for *De Novo* Assembly in Stickleback

We tested the performance of this partial-digest RAD-PE contig protocol by sequencing a fosmid from stickleback. After partially digesting the DNA using two

restriction enzymes with different 4 bp recognition sequences, NlaIII and Sau3AI, 1.0-5.0 kb DNA fragments were isolated prior to P1 ligation and shearing. Partially digested DNA samples were then carried through the RAD-PE contig protocol as described above.

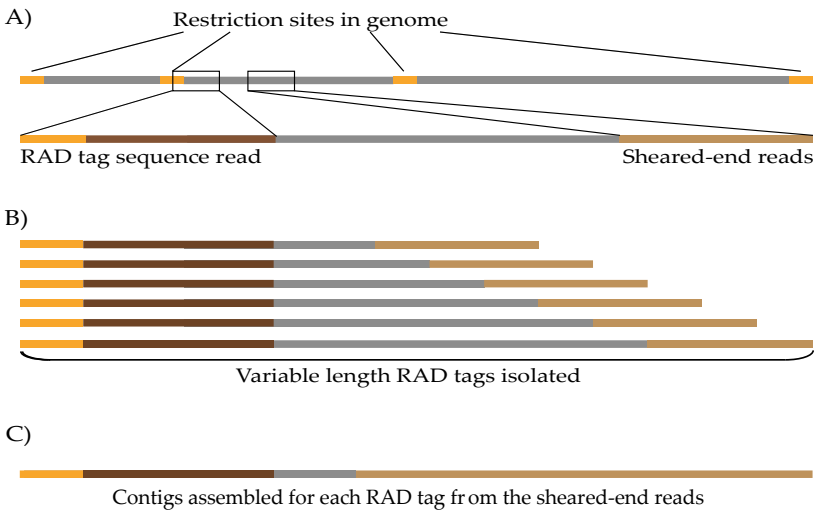

Figure 2.1. RAD paired-end contig libraries. (A) DNA fragments created by RAD tag library preparation have a restriction site (orange) at one end and a random sheared-end sequence (light brown) at the other. (B) Paired-end sequencing of RAD tag libraries allows the assembly of the sheared-end sequences from each RAD tag (dark brown) to be locally assembled into contigs on an individual basis (C). The distance at which the random end sequence lies, and hence the length of the contigs assembled, is dictated by the amount of shearing and the size of fragments isolated during the gel extraction step in the protocol.

Because the samples were over-sequenced (>3 million reads total), we removed reads that increased coverage over a 30x threshold, leaving 2 million reads for the assembly. We also tuned the Velvet parameters for each RAD site using a script that

tested three different word lengths and chose the assembly with the longest total contig lengths for that site (Figure 2.2B). The partial-digest strategy produced overlapping contigs as predicted (see Figure 2.2C) with an N50 length of 481 nucleotides; however, the assembled contigs mapped to two different regions of the genome, suggesting there were two fosmids present in the original prep.

Methods

Stickleback genomic DNA was isolated from pectoral fin clips using the DNeasy Tissue Kit (Qiagen). *E. coli* genomic DNA was acquired from the REL606 strain (provided by the Bohanan lab, UO) and from type B cells, ATCC 11303 strain (USB Corporation). Stickleback fosmids were isolated from genomic DNA using the CopyControl™ Fosmid Library Production Kit (Epicentre).

1.0 µg of genomic DNA from each individual (H2 -141, L2-110) was digested for 60 min at 37° C in a 50 µl reaction volume containing 5.0 µl 10x Buffer 4 and 10 units (U) SbfI-HF (New England Biolabs [NEB]). Samples were heat-inactivated for 20 min at 65° C. 4.0 µl of barcoded SbfI-P1 Adapter (100 nM), a modified Illumina© adapter (2006 Illumina, Inc., all rights reserved; top oligo: 5'-AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTxxxxxTGC*A-3' [xxxxx = barcode (AGAGT-H2; CAGTC-L2), * = phosphorothioate bond]; bottom oligo: 5'-Phos-xxxxxAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGGTCGCCGTATCAT*T-3'), was added to each sample along with 0.6 µl rATP (100mM,

Promega), 1.0 µl 10x NEB Buffer 4, 0.5 µl (1000 U) T4 DNA Ligase (high concentration, NEB), 3.9 µl H2O and incubated at room temperature (RT) for 30 min. Samples were again heat-inactivated for 20 min at 65° C, combined, and randomly sheared (Bioruptor) to an average size of 500 bp. The sheared sample was purified using a QIAquick Spin column (Qiagen) and run out on a 1.25% agarose (Sigma), 0.5x TBE gel.

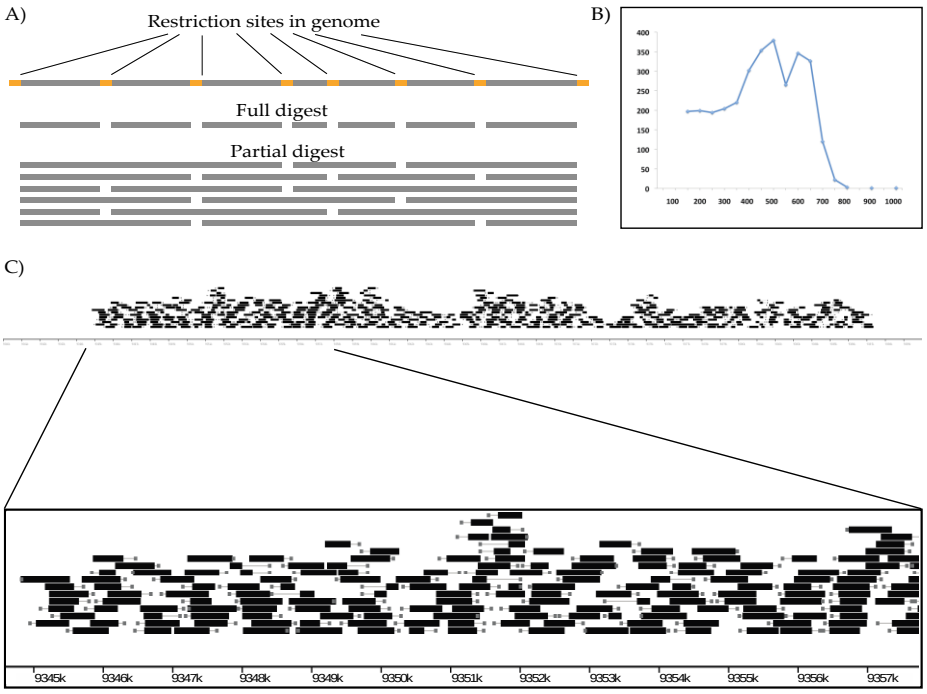

Figure 2.2. *De novo* assembly of a fosmid using Partial-digest RAD paired-end libraries in stickleback. (A) Incomplete digestion of DNA with a frequently cutting restriction enzyme creates overlapping restriction fragments. Preparing RAD-PE libraries from a stickleback fosmid following partial digestion with two frequent cutters resulted in contigs up to 1000 bp long and an N50 length of 481 nucleotides. (B) Shows the distribution of contigs built from the two libraries. (C) Mapping the contigs (black bars) from each RAD tag (grey boxes) back to the stickleback reference sequence demonstrated overlapping coverage over an ~40 kb stretch of the genome with a zoom on part of the assembly displayed below.

A smear of DNA approximately 300-800 bp was isolated with a clean razor blade and purified using the MinElute Gel Extraction Kit (Qiagen). The Quick Blunting Kit (NEB) was used to polish the ends of the DNA in a 25 µl reaction volume containing 2.5 µl 10x Blunting Buffer, 2.5 µl dNTP Mix and 1.0 µl Blunt Enzyme Mix. The sample was purified and incubated at 37° C for 30 min with 10 U Klenow Fragment (3′-5′ exo-, NEB) in a 50 µl reaction volume with 5.0 µl NEB Buffer 2 and 1.0 µl dATP (10 mM, Fermentas), to add 3′ adenine overhangs to the DNA. After another purification, 1.0 µl of Paired-End-P2 Adapter (PE-P2; 10 µM), a divergent modified Illumina© adapter (2006 Illumina, Inc., all rights reserved; top oligo: 5'-Phos-GATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGATCAGAACAA-3′, bottom oligo: 5′-CAAGCAGAAGACGGCATACGAGATCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATC*T-3′), was ligated to the DNA fragments at RT. The sample was purified and eluted in 50 µl. 25 µl of the eluate was digested again with SbfI for 30 min to remove rare genomic DNA concatemers formed from re-ligation of short fragments with two SbfI restriction sites within 500 bp. The sample was purified, eluted in 50 µl and quantified using the Quant-iT™ dsDNA HS Assay Kit and Qubit™ fluorometer (Invitrogen). ~40 ng was used as template in a 100 µl PCR amplification with 50 µl Phusion Master Mix (NEB) and 4.0 µl modified Illumina© amplification primer mix (10 µM, 2006 Illumina, Inc., all rights reserved; P1-forward primer: 5′-AATGATACGGCGACCACCGA-3′, P2-reverse primer: 5′-CAAGCAGAAGACGGCATACGA-3′). Phusion PCR settings followed product guidelines (NEB) for a total of 14 cycles with an annealing temperature

of 65° C. The library was cleaned through a column and gel purified, excising DNA ~350-850 bp in size in an inverted triangle shape. PCR amplification of a wide-range of fragment sizes often results in biased representation of amplified products with an increased number of short fragments. We found this to be true in our current protocol, but reduced the effects by selecting a triangular slice during gel extraction to reduce the level of short fragment lengths from the PCR reaction. The sample was diluted to 10 nM and sequenced on the Paired-end module of the Genome Analyzer II following Illumina protocols for 2x60 bp reads. Sequences are available at the NCBI Short Read Archive (http://www.ncbi.nlm.nih.gov/Traces/sra; accession number SRA024496.1).

Raw sequence reads were processed using custom Perl scripts (by Eric A. Johnson), to optimize read number and reduce artifacts within the data. Barcodes, if present, were trimmed from the raw reads. Reads with many poor quality scores were removed. The number of reads from each RAD site was tracked, and RAD sequences above a threshold were considered repetitive and removed. Single mismatch derivatives of these repetitive RAD sequences were also removed. RAD sites with a number of reads below a threshold were also removed from further analysis, as the associated paired-end reads would therefore lack sufficient coverage for calling SNPs or were likely to be sequence-error created artifacts.

The paired-end reads from each passing RAD site passing the above tests were sent to the Velvet assembler with a word length parameter that increased with increasing depth. Separate Velvet assemblies were also run with a fixed low and high word length, and the best assembly was chosen from the three trials based on the total assembled length of contigs. For the long insert assembly, the paired-end reads from each RAD site

were assembled with a word length of 41. The paired-end reads were re-assembled with a predicted optimal word length based on coverage and the first assembly contigs included as long read sequences to help guide the assembly at repeats.

SNP calling was performed by aligning the sequence reads from each individual to the assembled contigs with Novoalign. Mismatches were filtered to include only high quality nucleotides and tracked by sample. SNPs were called using a simple thresholding.

Multiple digestion reactions were set up for each DNA sample containing either 1.0 µg of each fosmid DNA sample (BP11.12H 7e2 sox9) or 2.0 µg of *E. coli* REL606 genomic DNA, 5.0 µl 10x Buffer 4, 100 µg/ml BSA and 2 U of NlaIII or Sau3AI (NEB). The reactions were incubated at 37° C in a 50 µl reaction volume for multiple lengths of time in order to achieve a spectrum of partially-digested to fully-digested DNA fragments. Digested samples were heat-inactivated for 20 min at 65° C and run out on a 1.0% agarose gel. A smear of DNA approximately 1.0-5.0 kb was isolated for each sample with a clean razor blade and purified. The isolated samples were quantified and the remaining DNA was ligated to enzyme-specific P1 Adapters (1.0 µM), modified Illumina© adapters (2006 Illumina, Inc., all rights reserved; NlaIII-P1 top oligo: 5'-AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTCATG-3'; NlaIII-P1 bottom oligo: 5'-Phos--AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGGTCGCCGTATCATT-3'; Sau3AI-P1 top oligo: 5'-AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT-3'; Sau3AI-P1 bottom oligo: 5'-Phos-GATCAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGGTCG

CCGTATCATT-3'), as described above, at a 10:1 molar ratio of adapter to DNA ends (assuming an average genomic DNA fragment length of 3.25 kb). Samples were heat-inactivated for 20 min at 65° C and randomly sheared to an average size of 500-800 bp. Sheared samples were purified, run out on a 1.0% gel and DNA smears 200-800 bp (200-1200 bp for the *E. coli* samples) were isolated and purified. DNA polishing and 3′ dA-overhang addition was carried out as described. PE-P2 ligations were carried out with 0.5 µl PE-P2 Adapter. Samples were purified, eluted in 50 µl and quantified. 20 ng of template was used in a 100 µl, 14-cycle Phusion PCR amplification with 25 µl Master Mix and 2.0 µl amplification primer mix. Libraries were cleaned and gel purified, excising DNA ~250-850 bp (250-1250 bp for the *E. coli* samples) in a triangle shape as above, diluted to 10 nM, and sequenced on the Paired-end module of the Genome Analyzer II following Illumina protocols for 40x80 bp reads.

Discussion

RAD tags are typically used to sample a portion of the genome, allowing high coverage at a desired number of loci. However, we modified the protocol to produce RAD-PE contigs that overlap over a genome using partial-digests with a frequent cutter. The advantage of this approach is that the short read sequences are first assembled into contigs, which can then be ordered into a genome-wide assembly. Whereas whole genome shotgun assemblies require specialized and expensive computational resources, RAD-PE contigs can be assembled on any computer.

The many challenges of whole genome assembly are mitigated by local assembly. Short read sequences have a high error rate, so for a whole genome assembly every sequence must be searched against all others using relaxed alignment parameters. But then related regions of the genome and repeats become indistinguishable. Also, sequences must have a long region of overlap to be pieced together in whole genome assemblies, as shorter words are found throughout the genome. When RAD-PE contigs are assembled, the small region size allows for easy alignment of even high-error sequences, and short regions of overlap are sufficient to piece sequences together.

Genome assembly programs like Velvet require the user to choose parameters such as word length and expected coverage. Even the best whole genome shotgun methods create peaks and valleys of coverage across a genome and the genome itself has regions of low and high complexity. Despite this variation, during assembly a median value for each parameter is chosen and the assembler therefore is less optimal in those regions that differ from that median. Our scripts collect the reads from a particular region and attempt to optimize the assembly for that single region by removing excessive reads and adjusting the indexing word length in response to the predicted coverage, with low coverage assemblies using a short word length, allowing sparse reads to join together and high coverage assemblies using a longer word length to bridge non-unique short sequences in the region. We also routinely tried a fixed low, fixed high and this predicted optimal word length for each region and evaluated the results to choose the best assembly for further use. Velvet can be recompiled to use longer word lengths than the default maximum of 31, but this greatly increases the memory requirements for an assembly. While this is a problem for whole genome shotgun assemblies that already require

hundreds of gigabytes of memory, we took this step for our assembly of the long-insert RAD-PE contigs without difficulty due to the low memory requirements of local assembly.

We showed the utility of using RAD-PE contigs for de novo assembly of large genomic regions by performing a partial-digest RAD-PE contig approach on a fosmid from stickleback using two different high-frequency cutters.

RAD paired-end contigs provide a low-cost method for SNP discovery in a format suitable for high-throughput genotyping platforms that require flanking sequence for primer design. It is possible to use platforms such as Roche 454 to achieve similar read lengths; however, accurate SNP discovery requires low error rates and sufficient depth of coverage to sample both chromosomes and determine heterozygosity. Although pricing of sequencing platforms rapidly change, a similar SNP discovery project using the 454 platform would currently cost more than ten times as much as RAD-PE contig sequencing. The 8 million reads used to create greater than 50,000 contigs and find more than 40,000 SNPs between the two stickleback samples are, at this time, just one quarter of a single Illumina GAIIx lane (1/28th of a run), whereas similar coverage would require at least one half of a full 454 run.

A related strategy to RAD paired-end contigs, termed subassembly, was recently described [3]. The complexity reduction step in subassembly is achieved randomly by dilution and amplification rather than restriction digestion, and subassemblies use the end sequence of the amplified fragments as an index rather than a restriction cut site sequence. As a result, subassembly does not create contigs at the same loci between

samples, making the several hundred nucleotide contigs it produces useful for de novo assembly rather than SNP discovery.

There is justified excitement over the next generation of sequencing platforms, which promise longer read lengths and simpler informatics. The longer assembly lengths created by long-insert RAD-PE contigs match the several kilobase output projected for the next generation of high-throughput sequencers, and the local assembly step also simplifies the computational needs of a de novo assembly project. While the next generation of sequencers currently suffer from a high error rate, RAD-PE contigs have a low error rate due to high coverage of any particular nucleotide. Therefore, users of high count, short read length sequencers can enjoy many of the benefits of long read lengths without the considerable expense of purchasing new systems and trouble of substantially altering their workflows.

Besides the short read lengths of current NGS sequencing platforms, the other major shortcoming of current DNA sequencing technology is the relatively high error rate of the DNA sequences generated with standard methods. This error rate is highly problematic when attempting to sequence genetically heterogeneous populations such as tumors or microbiomes to detect mutations present at below 1% of the population. To overcome these challenges, the PELE-Seq method was developed, which is described in Chapter III.

CHAPTER III

IDENTIFYING AND FILTERING NEXT-GEN SEQUENCING ERRORS

This work has been submitted to *BMC Genomics* 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)." The sequencing procedure and data analysis pipeline described in this chapter was developed by a number of lab members, including myself, Eric Johnson, and Ariel Royall. Paul Etter contributed substantially to this work by participating in the development of the method. I was the primary contributor to the optimization of the method and generated all the data. I developed the data analysis procedure and did all of the writing.

Introduction

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 [1, 2, 3]. 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 [4]. 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 [5,6,7,8].

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 [9]. PELE-Seq improves on the ORP method by incorporating dual-barcoding to filter out many types of PCR errors and library preparation artifacts, as well as a data analysis strategy that increases the specificity of SNP detection without a loss in sensitivity. 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 optimized to allow for the highly sensitive detection of rare polymorphisms 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 [10]. 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 low-frequency alleles that appear only in the laboratory adapted population.

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 3.1. 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, 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 [11]. 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.

Rare alleles are evaluated with LoFreq, which calls somatic variants using a Bonferroni-corrected P-value threshold of 0.05 [11]. 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 twice in two separate instances of high-sensitivity variant calling, once for each

barcode file. The final outcome of the PELE-Seq analysis is very high quality SNPs that have passed numerous quality control tests and filters.

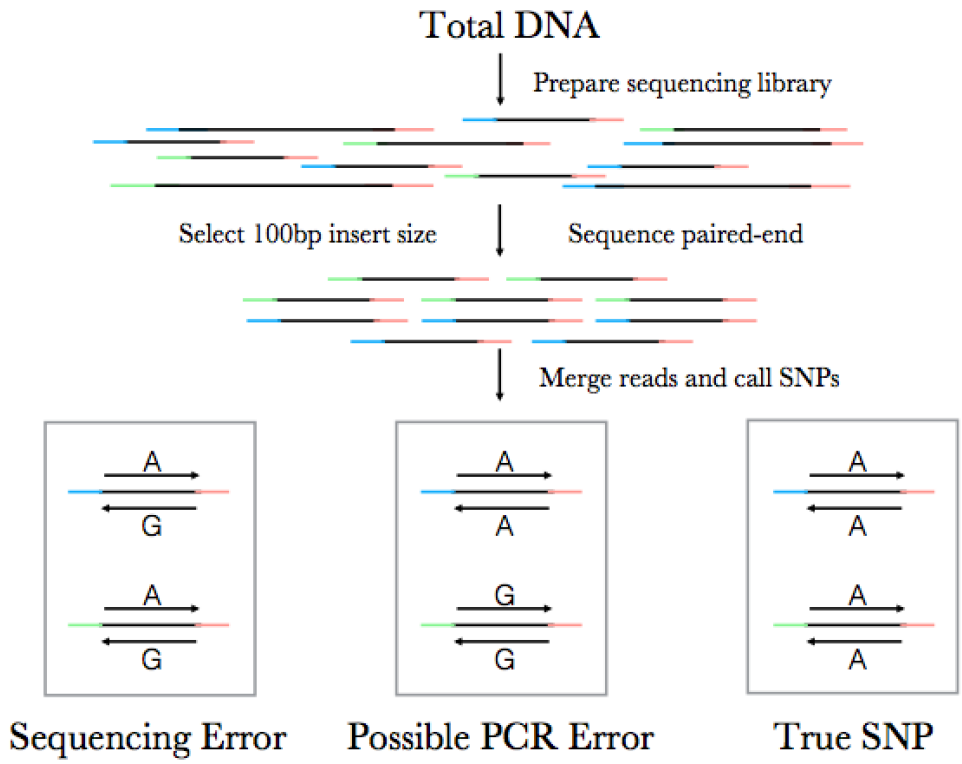

Figure 3.1. Overview of Paired-End Low Error Sequencing (PELE-Seq) library generation. DNA libraries with a 100bp insert size are paired-end sequenced using 100bp reads, generating an overlap region of approximately 100bp. The overlapping reads are merged into a consensus sequence and mismatching bases 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 Accuracy 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 [12,13]. 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, [14]. 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 3.1). 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).

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 3.2, 3.3). 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%.

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 3.2).

	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 3.1. Allele frequencies for known rare SNPs in control *E. coli* DNA mixtures labelled 1-4, 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.

PELE-Seq was 100% accurate at detecting rare alleles present at 0.3% with 30,000X read depth, compared to a 74% average accuracy level for standard Non-PELE Q30+ data. However, sequencing with ultra-high read depths (above 30,000X) resulted in the occurrence of false positive mutations in the PELE-Seq data, resulting in a 90% accuracy level, compared to 70% for standard DNA-Seq Q30+ data. The accuracy of standard DNA-Seq Q30+ data remained constant around 70%, regardless of the read depth used.

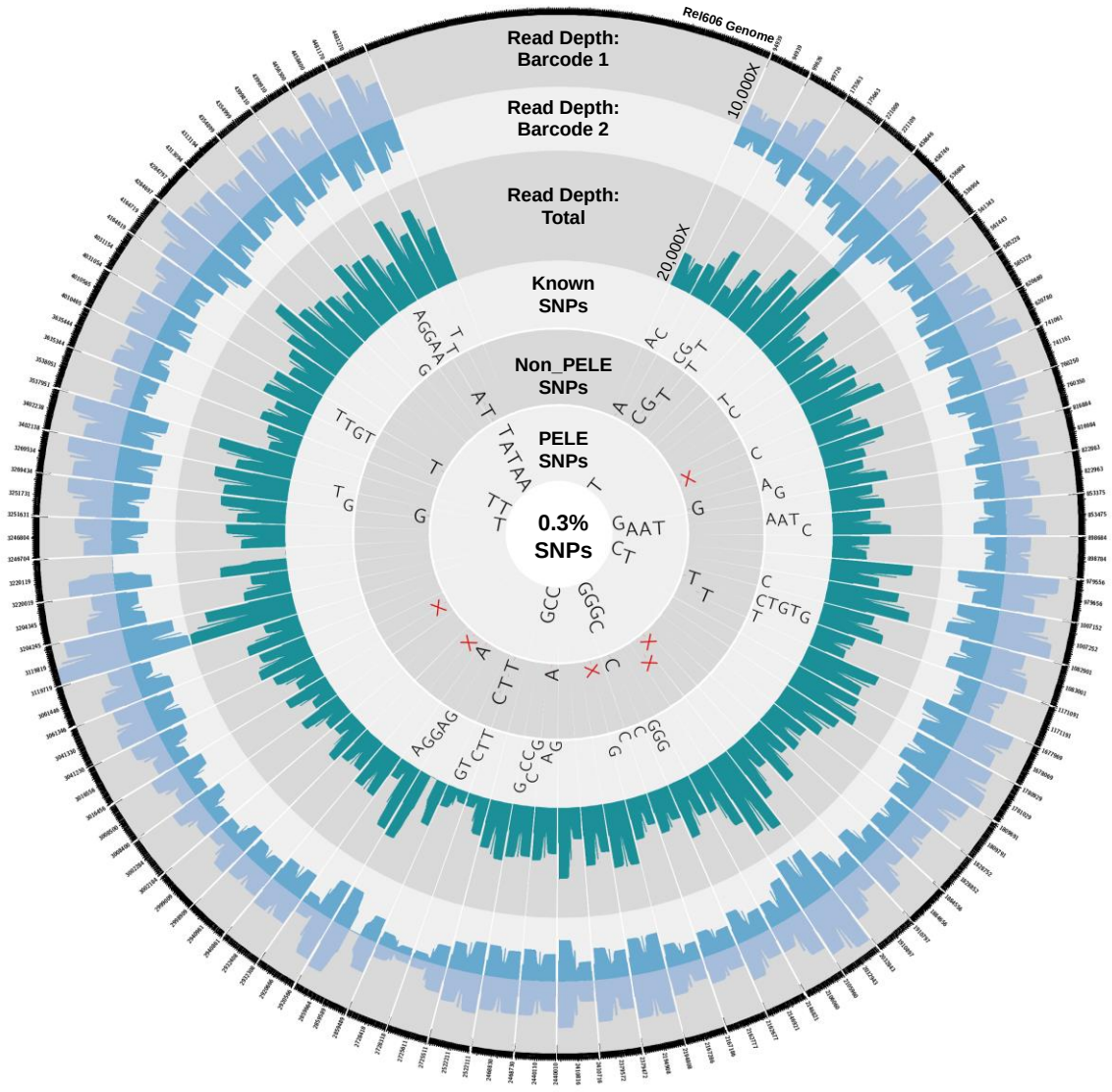

Figure 3.2. Detecting SNPs present at 0.3% frequency in *E. coli* control libraries with PELE-Seq and standard DNA-Seq methods at 20,000X average read depth. The read depths of the individual barcode files are plotted in blue, and the total read depth is plotted in green. The SNPs detected with PELE-Seq 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% accuracy, compared to 17 mutations and 70% accuracy achieved with non-PELE methods.

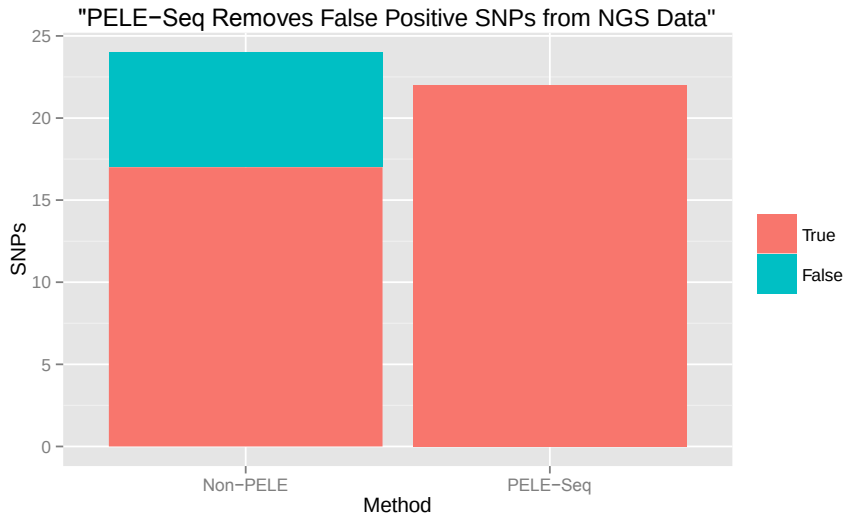

Figure 3.3. Sequencing a control *E. coli* DNA library containing 64 rare SNPs present at 0.3% allele frequency with PELE-Seq at 20,0000X read depth produces 100% accurate data, compared to 71% accuracy achieved with traditional sequencing methods. Traditional Non-PELE sequecning of the control libraries resulted in 7 false positive mutations, compared to zero with the PELE-Seq method.

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.

PELE-Seq Method

True Positives

ID	Position		Ref	Alt
1	175737	.	C	T
2	817018	.	A	G
3	853386	.	C	A
4	853407	.	G	A
5	853410	.	C	T
6	1007276	.	T	C
7	1083052	.	C	T
8	2146885	.	A	G
9	2146888	.	C	G
10	2146891	.	A	G
11	2162714	.	A	C
12	2468858	.	T	C
13	2468873	.	T	C
14	2468900	.	A	G
15	3269621	.	C	T
16	4010517	.	C	T
17	4010538	.	C	T
18	4399970	.	G	A
19	4399979	.	C	A
20	4458342	.	C	T
21	4458362	.	G	A
22	4458477	.	C	T

No False Positives

Accuracy: 100%

Standard DNA-Seq, Q30

True Positives

ID	Position		Ref	Alt
1	94900	.	T	A
2	175590	.	T	C
3	175596	.	A	G
4	175737	.	C	T
5	817018	.	A	G
6	1083055	.	C	T
7	1171239	.	C	T
8	2162714	.	A	C
9	2440136	.	G	A
10	2728328	.	C	T
11	2728331	.	A	T
12	2728367	.	A	C
13	2920697	.	G	A
14	3269492	.	A	G
15	4010580	.	C	T
16	4458342	.	C	T
17	4458362	.	G	A

False Positives

ID	Position		Ref	Alt
1	817075	.	G	A
2	2146912	.	T	C
3	2146924	.	G	A
4	2920673	.	A	G
5	2920679	.	G	C
6	2920763	.	A	G
7	3016457	.	C	T

Accuracy: 71%

ORP Method

True Positives

ID	Position		Ref	Alt
1	94966	.	T	C
2	175737	.	C	T
3	221029	.	C	T
4	741104	.	T	C
5	817018	.	A	G
6	853386	.	C	A
7	853407	.	G	A
8	853410	.	C	T
9	1007276	.	T	C
10	1083049	.	T	C
11	1083052	.	C	T
12	1083053	.	A	G
13	1083055	.	C	T
14	1083076	.	A	G
15	2146885	.	A	G
16	2146888	.	C	G
17	2146891	.	A	G
18	2162714	.	A	C
19	2440038	.	A	G
20	2468768	.	A	G
21	2468789	.	T	C
22	2468858	.	T	C
23	2468873	.	T	C
24	2728482	.	G	A
25	3269492	.	A	G
26	3269621	.	C	T
27	4010517	.	C	T

False Positives

ID	Position		Ref	Alt
1	458700	.	C	T
2	1884742	.	G	A
3	3251804	.	G	A
4	3402299	.	G	A
5	4010623	.	G	A
6	4031099	.	G	A

Accuracy: 82%

Table 3.2. Total SNP calls of 0.3% rare allele spike in libraries with PELE-Seq, DNA-Seq, and the ORP method. PELE-Seq data produces 100% accurate SNP calls, while standard DNA-Seq and the ORP method have accuracy rates of 71% and 82%, respectively.

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.

We identified several differences between the SNPs present in the wild nematodes compared to those found in the lab-adapted population (Figure 3.4). 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. 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 3.5). 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 3.3).

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 3.3. Fourteen SNPs present below 1% frequency in the wild *C. remanei* population increased in frequency at least 10x 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 43X 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 a gene predicted by the UCSC Genome Browser to be homologous to the *C. elegans* gene *ugt-5*, a UDP- Glucuronosyltransferase (Figure 3.6). The reads mapping to this SNP in the Integrative Genome Browser (IGV) are shown in Figure 3.7.

The lab-adapted worms also contained rare SNPs 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 3.8).

Methods

Wild isolates of *C. remanei* from Koffler Scientific Reserve at Jokers Hill, King City, Toronto, Ontario were graciously provided by Asher Cutter's lab (University of Toronto). "Isofemale strains" originating from 26 wild mating pairs were expanded to a population size of 2000 following the initial mating. All worms collected, and those in the experiment described below, were grown on nematode growth media (NGM) seeded with *E. coli* strain OP50. All collected strains were frozen within three generations of collection to minimize lab adaptation. To create a cohort representative of naturally segregating variation for experimental evolution, we thawed samples from each of the 26 isofemale strains and crossed them in a controlled fashion to promote equal contributions from all strains, including from mitochondrial genomes and Y chromosomes. The resulting genetically heterogeneous population was frozen after creation and was the ancestral population used for the experiment.

A lab-adaptation strain consisting of 1000-2000 mating individuals was propagated. The control populations were randomly culled to 1000 L1 larvae during each selective generation, for 23 generations. Each population was frozen ($N \geq 100{,}000$ individuals) periodically to retain a record of evolutionary change in the populations and

to ensure that worms did not lose the ability to survive freeze and thaw. Approximately 5000 individuals from the frozen populations were thawed to continue the evolution experiment, while the remaining 95,000 worms remained frozen for future phenotyping and genetic and genomic analyses. Populations were thawed for selection after a minimum of 24hrs at -80°C. Freezing occurred a total of 2 times during lab-adaptation selection. The lab-adapted population was also subjected to 11 rounds of bleaching/age-synchronization.

C. remanei genomic DNA was isolated using the DNeasy Tissue Kit (Qiagen). *E. coli* genomic DNA was acquired from REL606 strain (provided by the Bohannan lab, UO) and from W3110 strain (Life Technologies).

Restriction-Site Associated DNA (RAD) Sequencing was used to reduce the complexity of the *C. remanei* genome. For this application we used the restriction enzyme PacI, which has an AT-rich cut site. The complexity of the PacI RAD library was further reduced by digestion with NlaIII, which destroyed ~30% of the total RAD tags. The resulting PELE-PacI-RAD-Seq library was sequenced at 2000X coverage. RAD tags were present at approximately every 10kb throughout the genome.

Genomic DNA (2.0 μg) from each population was digested for 60 minutes at 37C in a 50 μL reaction volume containing 5.0 μL Buffer 1, 10 units (U) PacI (New England Biolabs [NEB]), and 0.5 μl 100X BSA (NEB). Samples were heat-inactivated for 20 min at 65 C. 1.0 μL of barcoded PacI-P1 adapter mixture (100 nM), a modified Illumina© adapter (2006 Illumina, Inc., all rights reserved; top oligo: 5'-ACACTCTTTCCCTACACGACGCTCTTCCGATCTxxxxx(xx)A*T -3'[xxxxx(xx) = barcode (TACGT, AGATCGA - ancestor; CTGCAA, GCTAGTC –evolved control), * =

phosphoro-thioate bond]; bottom oligo: 5'-Phos-xxxxx(xx)AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTG*T-3'), was added to each sample along with 0.6 ml rATP (100 mM, Promega), 1.0 µl 10X NEB Buffer 4, 0.5 µl (1000 U) T4 DNA Ligase (high concentration, NEB), 3.9 µl H2O and incubated at room temperature (RT) for 30 min.

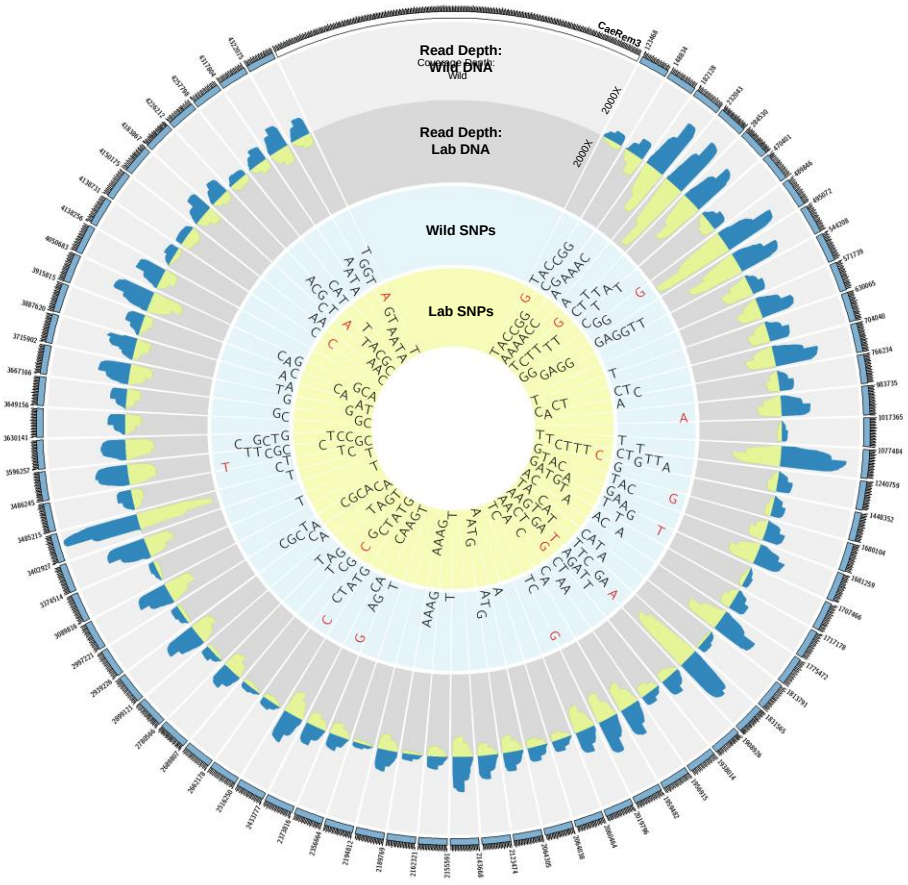

Figure 3.4. Total SNPs present in the wild and lab-adapted *C. remanei* populations. The inner yellow circle lists SNPs present in the lab-adapted population; the wild SNPs are listed in the blue circle. SNPs present in both the wild and lab-adapted populations are written with black letters. SNPs appearing in only the wild or lab-adapted populations are written with red letters.

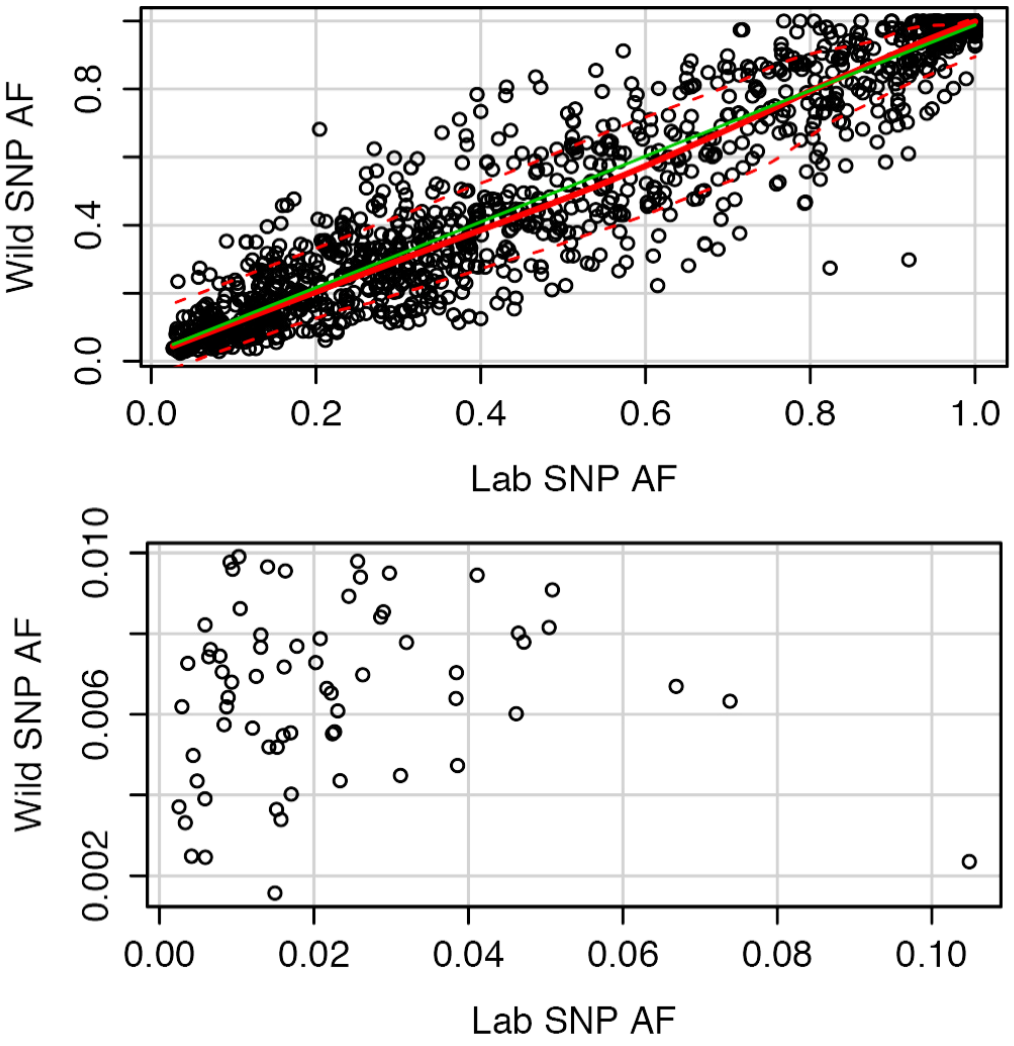

Figure 3.5. The allele frequencies of SNPs in the ancestral and lab-adapted populations of *C. remanei* worms. Each point represents a SNP in the genome. **Top)** Allele frequencies before and after lab-adaptation for all SNPs detected that are present in both populations. SNPs in the top left corner 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. Fourteen minor alleles present below 1% in the wild population increased in frequency at least tenfold after lab adaptation. Only SNPs present in both populations are plotted.

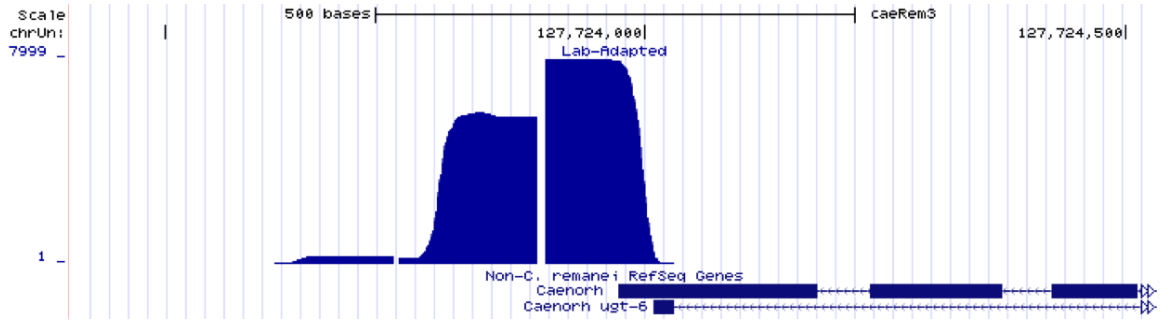

Figure 3.6. A RAD tag sequenced with PELE-Seq contains a SNP at position 127,723,967 of the caeRem3 (WUSTL) genome that maps to the predicted *C. elegans* gene *ugt-5* that was increased in frequency by 44X after 34 generations of lab-adaptation. The UGT pathway is a major pathway responsible for the removal of drugs, toxins, and foreign substances. *http://genome.ucsc.edu.*

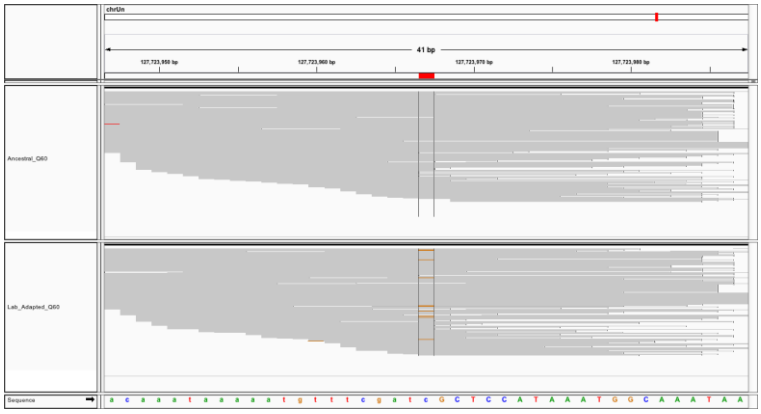

Figure 3.7. A SNP near the promoter region of *ugt-5* increases in frequency 43X after lab adaptation. A G>C transversion found at below 1% frequency in the ancestral *C. remanei* population has a 43X increase in frequency after 34 generations of laboratory adaptation. This SNP maps to the promoter region of predicted *C. elegans* gene *ugt-5*, which is an enzyme 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.

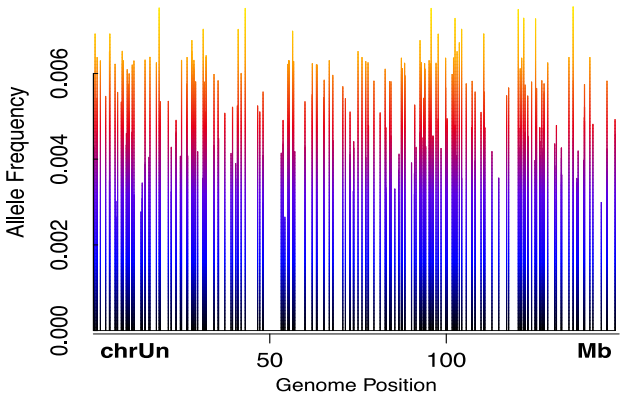

Figure 3.8. Allele frequencies and position of rare alleles detected only in the lab-adapted *C. remanei* population with PELE-Seq. Each vertical line represents a single *SNP*; the height of the line is proportional to the allele frequency. The detected SNPs had allele frequencies ranging from 0.0021 to 0.0075. The UCSC caeRem3 genome from WUSTL is composed of a single artificial chromosome named chrUn that is 146 megabases (Mb) long.

Samples were again heat-inactivated for 20 min at 65C, combined, and randomly sheared (Bioruptor) to an average size of 140 bp. The sheared sample was purified using a QIAquick Spin column (Qiagen) and run out on a 1.25% agarose (Sigma), 0.5X TBE gel. A tight band of DNA from 130-150 bp was isolated with a clean razor blade and purified using the MinElute Gel Extraction Kit (Qiagen). The Quick Blunting Kit (NEB) was used to blunt the ends of the DNA in a 25 µl reaction volume containing 2.5 µl 10X Blunting Buffer, 2.5 µl dNTP Mix and 1.0 µl Blunt Enzyme Mix. The sample was purified and incubated at 37C for 30 min with 10 U Klenow Fragment (3'-5' exo-, NEB) in a 50 µl reaction volume with 5.0 µl NEB Buffer 2 and 1.0 µl dATP (10 mM, Fermentas), to add 3' adenine overhangs to the DNA. After another purification, 1.0 ml of Paired-End-P2 Adapter (PE-P2; 10 mM), a divergent modified Illumina© adapter (2006 Illumina, Inc., all rights reserved; top oligo: 5'-Phos-

GATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGATCAGAACAA-3', bottom oligo: 5'-CAAGCAGAAGACGGCATACGAGATCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATC*T-3'), was ligated to the DNA fragments at RT. The sample was purified and eluted in 50 µl. The eluate was digested again with NlaIII to reduce library complexity. The sample was column purified and eluted in 10 µl. Two separate PCR amplifications were performed with each sample, each using 5µl of eluate as template, in a 50 µl volume with 25 µl Phusion Master Mix (NEB) and 1.0 µl modified Illumina© amplification primer mix (10 mM, 2006 Illumina, Inc., all rights reserved; P1-forward primer: 5' AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATC*T 3', P2-reverse primer: 5' CAAGCAGAAGACGGCATACG*A 3'). Phusion PCR settings followed product guidelines (NEB) for a total of 17 cycles with an annealing temperature of 65C. The libraries were pooled and cleaned through a column and gel purified, excising a tight band of DNA of 240 bp size. The sample was diluted to 1 nM and sequenced on the Paired-end module of the Genome Analyzer II following Illumina protocols for 100 bp reads.

Serial dilution of *E. coli* W3110 DNA with *E. coli* Rel606 DNA was performed to generate spike-in libraries with dilution levels ranging from 1:100 to 1:5000, at a concentration of 0.8 ng/µl. All dilutions were concentrated with a SpeedVac to 40 µl. 300 ng of genomic DNA from each dilution was digested for 60 minutes at 37C in a 50 µL reaction volume containing 5.0 µL Buffer 4, 10 units (U) SbfI-HF (New England Biolabs [NEB]). Samples were heat-inactivated for 20 min at 65 C. 2.0 µL of barcoded SbfI-P1

adapter mixture (100 nM), a modified Illumina© adapter (2006 Illumina, Inc., all rights reserved; top oligo: 5'-Phos-AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTxxxxxxTGC*A 3'[xxxxxx = barcode (mixture of two barcodes per sample), * = phosphoro-thioate bond]; bottom oligo: 5'-Phos-xxxxxxAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGGTCGCCGTATCAT*T-3'), was added to each sample along with 0.6 ml rATP (100 mM, Promega), 1.0 µl 10X NEB Buffer 4, 0.5 µl (1000 U) T4 DNA Ligase (high concentration, NEB), 3.9 µl H2O and incubated at room temperature (RT) for 30 min. Samples were again heat-inactivated for 20 min at 65C, combined, and randomly sheared (Bioruptor) to an average size of 140 bp. The sheared sample was purified using Agencourt AMPure XP beads at a 1X volume. The Quick Blunting Kit (NEB) was used to blunt the ends of the DNA in a 50 µl reaction volume, and the sample was purified using Agencourt AMPure XP beads at a 1X volume. The sample was incubated at 37C for 30 min with 10 U Klenow Fragment (3'-5' exo-, NEB) in a 50 µl reaction volume with 5.0 µl NEB Buffer 2 and 1.0 µl dATP (10 mM, Fermentas), to add 3' adenine overhangs to the DNA. After another 1X bead purification, 1.0 ml of Paired-End-P2 Adapter (PE-P2; 10 mM), a divergent modified Illumina© adapter (2006 Illumina, Inc., all rights reserved; top oligo: 5'-Phos-GATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGATCAGAACAA-3', bottom oligo: 5'-CAAGCAGAAGACGGCATACGAGATCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATC*T-3'), was ligated to the DNA fragments at RT. The sample was purified

and eluted in 40 µl. Ten separate PCR amplifications were performed with the sample, each using 4µl of eluate as template, in a 50 µl volume with 25 µl Phusion Master Mix (NEB) and 1.0 µl modified Illumina© amplification primer mix (10 mM, 2006 Illumina, Inc., all rights reserved; P1-forward primer: 5' AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCC GATC*T 3', P2-reverse primer: 5' CAAGCAGAAGACGGCATACG*A 3'). Phusion PCR settings followed product guidelines (NEB) for a total of 18 cycles with an annealing temperature of 65C. The libraries were pooled, cleaned through a QIAquick Spin column (Qiagen), and size selected with a Pippin Prep (Sage), collecting a tight band of DNA of 240 bp size. The sample was diluted to 1 nM and sequenced on the Paired-end module of an Illumina HiSeq 2500 following Illumina protocols for 100 bp reads.

Discussion

Current genomic studies of genetically heterogeneous samples, such as growing tumors acquiring *de novo* mutations, 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 [9]. 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 ultra-rare alleles that are only detectable after 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 43X 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 [15]. 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 [16]. 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.

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 43X 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.

In addition to sequencing rare alleles in a mixed population of individual organisms such as nematodes, PELE-Seq is useful for detecting *de novo* mutations in genetically heterogeneous environments such as tumors. The detection of rare mutations in a tumor is critical for an understanding of early tumorigenesis and tumor evolution. Sequencing tumors with standard NGS methods produces data containing an overwhelming number of false positive mutations, which cannot be distinguished from true mutations. PELE-Seq can filter out the false positive mutations in tumor sequencing data, and accurately identifying rare mutations. In Chapter IV, PELE-Seq is applied to detect rare mutations in the blood DNA of a human with disseminated osteosarcoma.

CHAPTER IV

HIGH-SPECIFICITY TUMOR SEQUENCING

Tumors are highly heterogeneous communities of cells that evolve through the accumulation of rare *de novo* mutations. Recently it has been shown that advanced tumors can actually contain several different types of genetically distinct cancer cells, each with a different prognosis and metastatic potential [1,2,3,4,5]. Tumor heterogeneity allows for tumor plasticity, and is a huge barrier preventing personalized cancer medicine. Because of tumor heterogeneity, a single biopsy often gives inaccurate and incomplete information regarding the genetic potential of a tumor. Often, the most aggressive cells in a tumor may represent only a fraction of the solid tumor and may not be detected in a standard single biopsy [6,7]. In addition, small and early tumors are difficult to extract and purify from nontumor cells, making their DNA challenging to detect. Because standard NGS methods cannot accurately detect rare mutations, we lack a clear understanding of the early tumor cascade of mutations that would represent the most attractive drug targets. A high-resolution systems-level understanding of the mutations involved in cancer progression is required. The PELE-Seq method was developed to improve the accuracy of rare mutation detection, and is described in Chapter III. As described below in Chapter IV, PELE-Seq was applied to sequence the blood DNA of an anonymous patient with disseminated metastases, and to identify the clinically-relevant mutations in the genome. Cancer patients often contain circulating tumor cells and cell-free DNA in their blood [8,9,10], and PELE-Seq can be used to accurately detect rare genetic variants found in the blood of a cancer patient.

PELE-Sequencing of a Human Osteosarcoma Blood Biopsy

Under many circumstances, biopsies cannot be done due to the location and nature of a tumor. Additionally, biopsies give incomplete information when a tumor is large and heterogeneous, and the driver cells of the tumor or metastases are rare compared to the overall tumor. Cancer DNA from disseminated metastases can exist in the blood, lymph, or cerebrospinal fluid at low frequencies, and can be detected with a liquid biopsy. However, the rare mutations are often challenging to detect due to the high error rate of standard NGS techniques, as described in Chapter III.

We applied PELE-Seq to sequence a whole-exome capture of blood DNA from an anonymous human patient with disseminated metastatic sarcoma, in collaboration with the Spellman lab at OHSU. We performed a PELE Sequencing and analysis, similar to that described in Chapter III, but without barcode information. The blood tumor DNA was sequenced to an average depth of coverage of 1000X, and screened for relevant mutations that are previously known to impact protein function in known cancer genes, based on evolutionary conservation of the affected amino acid in protein homologs [11,12].

We identified ~5000 rare mutations passing quality filters that were present in disseminated osteosarcoma blood DNA. Of the mutations passing quality filters, 36 were functionally relevant mutations in known cancer genes (Table 4.1), including a mutation in PIK3CA present at 1.2%. Several other interesting mutations were also present in the blood DNA, including mutations in Atrx and Lrp1, which were present at around 0.1%.

Discussion

Several of the mutations identified in the tumor blood sequencing are present in genes known to be involved in cancer initiation and progression. For example, Atrx is a known tumor suppressor [13], and Lrp1 is a known oncogene [14]. Unfortunately, none of the mutations discovered in the blood DNA of the anonymous osteosarcoma patient are currently targetable with cancer therapy drugs. Much work remains to determine the functional significance of the mutations discovered during this and other tumor sequencing projects. Cancer biologists have barely begun to understand the functional significance of the genetic variation of tumors, especially with regard to noncoding variants, including enhancer elements and noncoding RNAs [15,16]. The PELE-Seq method is a very useful and necessary first step in assessing and tracking the mutational landscape of the tumor environment. With PELE-Seq it is now possible to track the accumulation of mutations in a tumor, starting from early tumorigenesis, in order to identify the "driver mutations" responsible for tumor growth. The driver mutations of a tumor are very attractive drug targets because they are acquired early in tumor development and are probably necessary for the survival of the entire tumor [17]. Once we have a better understanding of the identity of all possible driver mutations for a tumor, one possible strategy involves treating a driver mutation before it is even detected in a tumor, which is the so-called "never mutation" approach [18]. Additionally, so-called "passenger mutations," which are random *de novo* mutations that have accumulated over time due to the high mutation rate of tumors, also play a role in modifying tumor

progression and are important to track and understand as well [19]. PELE-Seq can be applied to understand the entire spectrum of mutations found in a tumor over time.

PELE-Sequencing of blood biopsies and cell-free tumor DNA sequencing methods hold promise for future research for a basic understanding of tumor biology, as well as for clinical applications such as early detection and screening for relapse of recurrent metastases. Because the blood of a cancer patient with disseminated metastases contains a collection of DNA from various metastatic sites, sequencing the blood DNA would be very useful for detecting rare cancer mutations involved in tumor progression. PELE-Seq could also be very useful to predict the outcome of a drug therapy and to gauge the potential of tumor drug resistance, because it allows researchers to accurately assess the genetic potential of a tumor. In summary, PELE-Seq is an extremely useful method to understand tumor progression and metastases, and has many important applications in basic and clinical research.

In addition to small *de novo* mutations, tumors often contain several types of genomic alterations, including large-scale amplifications, deletions, and rearrangements. Chapter V describes work done to identify large-scale rearrangements in the genome of advanced mouse glioblastoma tumors.

ID	Mutation	Frequency	Percent	AA	Gene	Impact
10	hg19,10,7285549,T,C	(2/489)	0.408%	K364R	SFMBT2	medium
431	hg19,10,129141928,C,T	(3/1396)	0.214%	H1027Y	DOCK1	medium
474	hg19,11,3733913,C,T	(2/657)	0.304%	E892K	NUP98	medium
594	hg19,11,62294181,G,T	(2/1209)	0.165%	P2570T	AHNAK	high
668	hg19,11,118376088,C,T	(2/786)	0.254%	H3158Y	MLL	medium
830	hg19,12,57593033,C,A	(5/632)	0.791%	L3239M	LRP1	medium
981	hg19,13,49033921,C,A	(2/412)	0.485%	H686Q	RB1	low
1045	hg19,14,33293374,G,A	(2/517)	0.386%	D2119N	AKAP6	medium
1125	hg19,14,102483751,C,T	(2/406)	0.492%	S2696L	DYNC1H1	medium
1622	hg19,17,7416859,C,T	(2/3570)	0.056%	S1759L	POLR2A	medium
1626	hg19,17,7736178,G,A	(2/1004)	0.199%	V4304M	DNAH2	medium
1943	hg19,19,9082895,C,A	(2/975)	0.205%	E2975*	MUC16	stop_gain
2148	hg19,1,11318590,C,T	(2/486)	0.411%	V75I	MTOR	low
2216	hg19,1,26620746,G,A	(2/242)	0.826%	S170L	UBXN11	medium
2285	hg19,1,82402494,G,A	(2/504)	0.396%	E124K	LPHN2	medium
2299	hg19,1,94643665,C,T	(2/660)	0.303%	G847R	ARHGAP29	high
2530	hg19,1,152280899,C,T	(2/1912)	0.104%	E2155K	FLG	medium
2674	hg19,1,237886513,C,T	(2/1069)	0.187%	T3547M	RYR2	low
2944	hg19,2,47705556,G,A	(2/739)	0.27%	E786K	MSH2	high
3249	hg19,2,196636467,C,T	(2/922)	0.216%	V3784I	DNAH7	medium
3441	hg19,3,89259316,C,T	(2/1069)	0.187%	Q154*	EPHA3	stop_gain
3443	hg19,3,89457251,G,A	(2/945)	0.211%	E578K	EPHA3	low
3838	hg19,4,126370200,G,A	(2/551)	0.362%	E2677K	FAT4	low
3902	hg19,5,13794112,G,A	(4/383)	1.044%	A2648V	DNAH5	low
4019	hg19,5,89979622,C,T	(2/963)	0.207%	H1962Y	GPR98	low
4590	hg19,7,140500235,C,T	(2/395)	0.506%	E303K	BRAF	low
3378	hg19,3,38760206,C,A	(3/488)	0.614%	A1207S	SCN10A	medium
3547	hg19,3,178916876,G,A	(9/719)	1.251%	R88Q	PIK3CA	medium
4302	hg19,6,160952811,G,A	(2/494)	0.404%	R4533C	LPA	high
4598	hg19,7,142568336,C,T	(2/376)	0.531%	S952L	EPHB6	medium
4614	hg19,7,151927031,G,A	(2/3214)	0.062%	H985Y	MLL3	medium
4639	hg19,8,14095115,C,T	(3/2982)	0.1%	G124E	SGCZ	medium
4682	hg19,8,48689506,C,T	(2/998)	0.2%	W4027*	PRKDC	stop_gain
4684	hg19,8,48746823,C,T	(2/1406)	0.142%	V2696M	PRKDC	medium
4721	hg19,8,100514023,C,T	(2/772)	0.259%	Q1327*	VPS13B	stop_gain
5021	hg19,X,76907752,G,A	(2/1942)	0.102%	S1470F	ATRX	medium

Table 4.1. Clinically relevant mutations detected in the blood DNA of an anonymous human patient with disseminated osteosarcoma, in collaboration with the Spellman lab at OHSU. Of the 5000+ rare variants identified in the blood DNA, 36 were predicted to be functionally relevant mutations affecting protein function in known cancer genes. The mutations detected are very rare compared to the majority of the blood DNA, with allele frequencies ranging from 0.1%-1%.

CHAPTER V

MOUSE GLIOBLASTOMA GENOME SEQUENCING

Cells must accumulate many mutations in order to acquire the hallmarks of cancer. In order to understand the evolutionary process of genome adaptation that these tumors undergo during malignant transformation, we must know which genes are mutated or altered during transformation. It is possible that there is a predictable sequence events leading to malignant transformation, and that certain early mutations are required for tumorigenesis. These early mutations, referred to as "driver mutations," are very promising drug targets, as described in Chapter IV. Once identified, driver mutations can be screened for and treated in a rational manner, through the development of targeted drug therapies. Besides point mutations, tumors often contain many types of genetic alterations including small insertions and deletions, and large-scale alterations and rearrangements. Large scale rearrangements are difficult to detect with Next-Generation Sequencing techniques, due to the short reads produced by the sequencers, the high error rate of the nucleotide sequence, and the high depth of coverage necessary to obtain the statistical power necessary to call a rearrangement [1,2].

The Mosaic Analysis with Double Markers (MADM) system in mice is a very useful tool for investigating early tumorigenesis and holds promise for the discovery of driver mutations [3]. MADM mice are engineered to lack tumor suppressor genes in a mosaic manner that resembles a natural loss-of-heterozygosity event in a human tumor. The tumors produced with the MADM system are superior to standard tumor suppressor knockout mice because the MADM system generates rare clonal tumors whose origins

more closely resemble human tumors. In addition, the MADM tumors can be identified by GFP expression. The GFP-labeled tumors can be dissected and studied with various techniques including DNA-Seq, RNA-Seq, immunohistochemistry, and flow cytometry. The work below describes the whole genome sequencing (WGS) of three such tumors dissected from mice engineered to lack functional p53 and Nf1genes in a small fraction of their neuroblasts, referred to as p53/Nf1 knockout (KO) MADM mice.

Mosaic Analysis with Double Markers (MADM) Mouse Model of Glioblastoma

The Mosaic Analysis with Double Markers (MADM) system allows for the creation of GFP-labeled mosaic tumors through simultaneous tumor suppressor deletion and GFP expression in rare mutant cells, generating realistic tumors that can be easily identified and studied. Because individual mutant cells are labeled, the MADM system allows for tumors to be analyzed *in vivo* at single-cell resolution. The mosaic double knockout of p53 and Nf1 has previously been shown to be an effective method of inducing glioblastomas in mice with high penetrance [3]. The p53/Nf1 KO MADM mice are engineered to contain the MADM p53/Nf1 cassette on chromosome 11. Upon tamoxifen injection, Cre is expressed in neuroblasts, leading to a small subset of cells undergoing sporadic inter-chromosomal mitotic recombination. The genomic rearrangement causes loss of heterozygosity (LOH) in the p53 and Nf1 genes, simultaneously with labeling of mutant cells, to generate uniquely identifiable homozygous mutant cells in a heterozygous background. These mice develop

glioblastoma with 100% penetrance and are an ideal mouse model for identifying early driver genes of glioblastoma.

The MADM model has several advantages over traditional cancer models based on tumor suppressor deletion, allowing for more precise control of the in vivo tumor environment. GFP-labeling of tumor cells allows for cell lineage tracing throughout the metastatic process, while RFP-labeling of wild-type cells provides an internal control. Because MADM provides single-cell resolution, the initiation of tumorigenesis can be directly analyzed, making MADM an ideal system for discovering driver mutations of glioblastoma.

In the p53/Kf1 KO MADM glioma model, p53 and Nf1 knockout alone is insufficient to cause malignant transformation. Typically only a small fraction of green (GFP-labeled) p53/Nf1 KO cells undergo malignant transformation. This observation provides support for the multi-hit hypothesis, which states that a cell must acquire several significant mutations for tumorigenesis to occur. Work in the Zong lab has shown that p53/Nf1 MADM knockout mouse neuroblasts do not transform into glioma cells until after they have become oligodendrocyte precursor cells (OPCs), implying that an important developmental step occurs in OPCs leading to glioma progression [4,5]. We sought to identify potential driver mutations of glioblastoma by sequencing advanced p53/Nf1 KO mouse gliomas and searching for mutations involved in OPC development and/or cell cycle regulation.

Mouse Glioblastoma DNA Sequencing

The whole genomes of three gliomas from p53/Nf1 knockout MADM mice were sequenced at 4-6x average coverage with standard DNA-Seq, and aligned to the mm9 mouse genome. Two of the tumors had matched nontumor samples that were also sequenced at 4-6x coverage (tumors 20641 and 20648). As expected, the tumor samples contained very few reads mapping to the deleted regions of the p53 and Nf1 genes. Sequencing reads mapping to exons 2-5 of the trp53 gene displayed on the Integrative Genomics Viewer (IGV) [6,7] are shown in Figure 5.1. The few wild type reads present in the knockout region are most likely due to immune and vessel cells in the tumor.

Thousands of single nucleotide variants were identified in the tumor samples that were less frequent in the matched nontumor samples. The variants were assessed for predicted impact on protein function by the program SnpEff [8]. In tumor 20648, 16 mutations were predicted to have a moderate functional impact in protein (Table 5.1). In tumor 20641, 18 mutations were predicted to have a moderate functional impact in protein (Table 5.2). However, none of the detected mutations in either tumor were known to be involved in OPC development, the cell-cycle, or any other frequently observed cancer processes. In addition, because of the low depth of coverage of the sequencing, it could not be ruled out that some variants may be strain-specific sequence variants that did not appear in the nontumor sequences due to statistical chance. Due to the lack of relevant point mutations discovered, the focus was switched to instead investigate large-scale genome rearrangements.

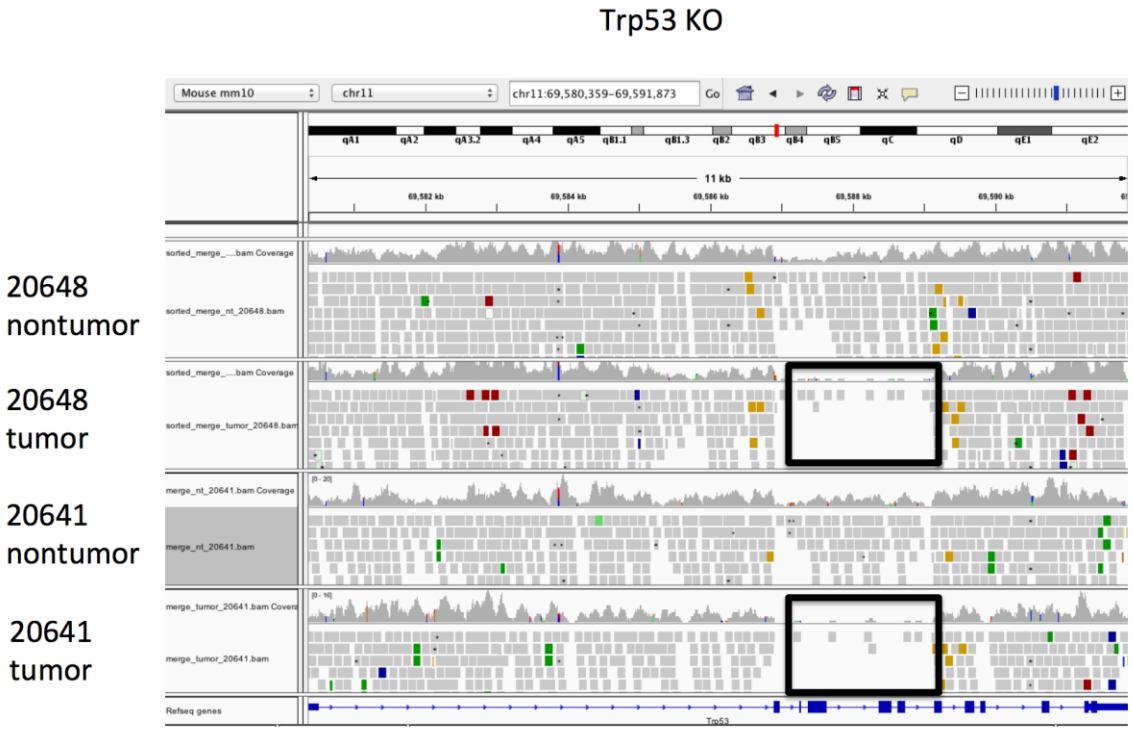

Figure 5.1. Sequencing reads mapping to the p53 knockout regions of the glioblastoma MADM cassette on chromosome 11 of mouse genome mm10 for tumors 20641 and 20648, and matched nontumor samples, using the Integrative Genome Viewer (IGV). The genomes of p53/Nf1 knockout MADM mice lack exons 2-4 of the gene Trp53 (p53). This is one of two specific regions lost due to homologous recombination of the MADM cassette in mouse neuroblasts.

By plotting the coverage depth of the tumor genomes as a bigwig file on the UCSC genome browser [9,10], and scrolling through the genome in 10 megabase (Mb) sections, four large gene amplifications were identified in the three glioblastoma tumors. These amplifications were absent in the matched nontumor samples, which had low read coverage throughout the genome. Each tumor contained one or two large genome amplifications around a well-known cancer gene, as described below. Besides these four

Gene	Chr	Position	Ref	Alt	Nontumor AF	Tumor AF	Type
Abca4	chr3	122069065	C	T	0.12	0.50	Missense
Abca4	chr3	122121788	G	A	0.08	0.57	Retained intron
Bcl2l13	chr6	120862938	A	T	0.11	0.50	Missense
Chl1	chr6	103711153	T	G	0.00	0.29	Missense
Col6a4	chr9	106062889	C	T	0.08	0.33	Missense
Furin	chr7	80391877	T	C	0.14	0.61	Retained intron
Itga6	chr2	71840880	C	T	0.11	0.34	Missense
Itga6	chr2	71855890	A	G	0.09	0.29	Missense
Lamc2	chr1	153141663	T	C	0.11	0.57	Missense
Lrrfip1	chr1	91115402	G	A	0.10	0.46	Missense
Muc4	chr16	32752234	G	C	0.09	0.50	Missense
Muc4	chr16	32753901	C	A	0.00	0.27	Missense
Muc4	chr16	32753927	A	T	0.06	0.30	Missense
Nanog	chr6	122707846	A	G	0.11	0.46	Missense
Rac1	chr5	143507067	G	A	0.04	0.23	Nonsense

Table 5.1. Tumor 20648 Mutations and allele frequencies in the nontumor and tumor genomes, aligned to mouse mm10 genome.

Gene	Chr	Position	Ref	Alt	Nontumor AF	Tumor AF	Type
Akna	chr4	63371838	G	A	0.10	0.44	Missense
Atp1a4	chr1	172255054	T	C	0.14	0.46	Missense
Atp5e	chr2	174462517	A	C	0.12	0.29	Retained intron
Cdhr1	chr14	37079410	T	C	0.14	0.42	Missense
Cnot8	chr11	58104794	GG	G	0.00	0.50	Splice-site donor
Gtse1	chr15	85862441	T	C	0.12	0.42	Missense
Gucy2d	chr7	98459044	G	A	0.10	0.44	Missense
Hnrnpab	chr11	51601417	CTTGTA	C	0.00	0.54	Frameshift
Lig1	chr7	13308507	C	A	0.00	0.58	Missense
Lrrfip1	chr1	91115879	A	G	0.11	0.57	Missense
Muc4	chr16	32752825	C	T	0.08	0.50	Missense
Muc4	chr16	32756159	C	A	0.11	0.50	Missense
Nop2	chr6	125144447	G	A	0.17	0.53	Missense
Pramef6	chr4	143895595	T	A	0.08	0.29	Missense
Pramef17	chr4	143991586	G	GA	0.00	0.50	Frameshift
Ran	chr5	129022093	A	G	0.00	0.19	Missense
Stab1	chr14	31150240	C	A	0.08	0.50	Missense
Ush2a	chr1	188576187	T	C	0.17	0.40	Missnese

Table 5.2. Tumor 20641 Mutations and allele frequencies in the nontumor and tumor genomes, aligned to mouse mm10 genome.

large amplifications and the expected p53/Nf1 knockout, the tumor genomes were completely intact, with smooth and even coverage throughout the genome. Smaller

amplifications and deletions were present throughout the tumor genomes, but those were present in both the tumor and the matched nontumor samples.

In tumor 20648, the smaller of the two tumors with matched nontumor tissue, a large, 400 kilobase (kb) amplification was discovered around the genes Ephb2 and Ephb8 on chromosome 4 of the mouse genome, shown in Figure 5.2. Ephrins are a class of receptor tyrosine-kinases, cell-surface molecules involved in juxtacrine signaling. Ephb2 is an ephrin receptor known to regulate glioma cell invasion [11,12]. Ephb2 and Ephb8 are not normally expressed in OPCs. Aside from the large amplification around Ephb2, there are no other large amplifications or deletions in the entire genome.

In tumor 20641, which was larger than tumor 20648, there were two separate large-scale amplifications, shown in Figures 5.3 and 5.4. Around the known glioblastoma oncogene MycN on chromosome 12, there was a 1.5 Mb amplification. There was also a 6 Mb amplification around the gene Sox4 on chromosome 13, which is an important developmental gene for OPCs [13,14,15]. As in tumor 20648, aside from the two large amplifications, the tumor genomes are very intact and uniform, with even coverage throughout the genome.

In the final tumor analyzed, which was sequenced previously without a matched nontumor sample, a 1 Mb amplification was discovered on chromosome 7. This genome amplification contained several genes, including Cd22, Nfkbid, and two Cox genes (Figure 5.5).

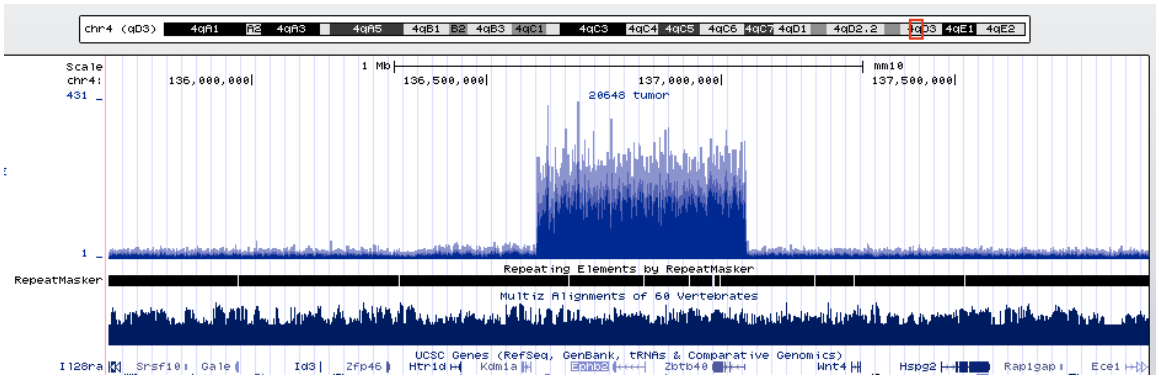

Figure 5.2. Sequencing reads that map to a 400kb region in chromosome 4 of mouse genome mm10 are ten times more abundant than reads mapping to other regions of the genome in tumor 20648. This 400kb region contains at least 5 genes, including Ephb2 and Ephb8. *http://genome.ucsc.edu.*

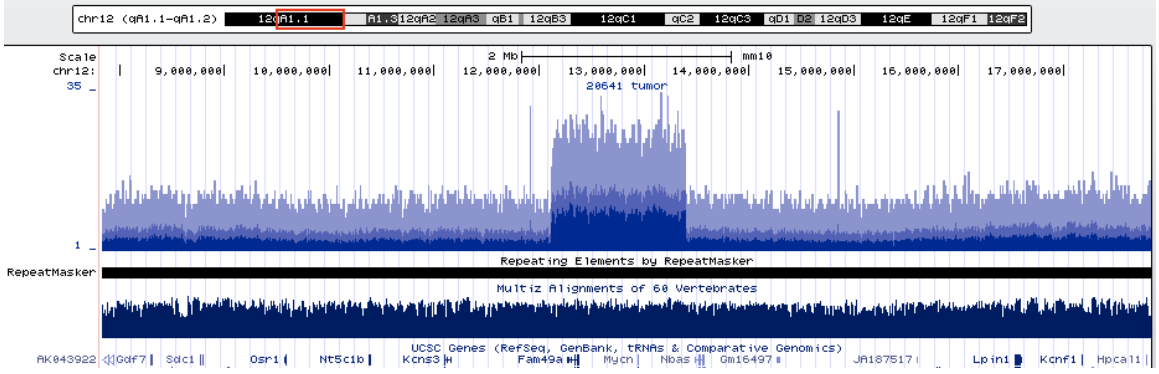

Figure 5.3. Sequencing reads that map to a 1.5Mb region in chromosome 12 of mouse genome mm10 are four times more abundant than reads mapping to other regions of the genome in tumor 20641. This amplified region contains the gene MycN, a known glioblastoma oncogene. *http://genome.ucsc.edu.*

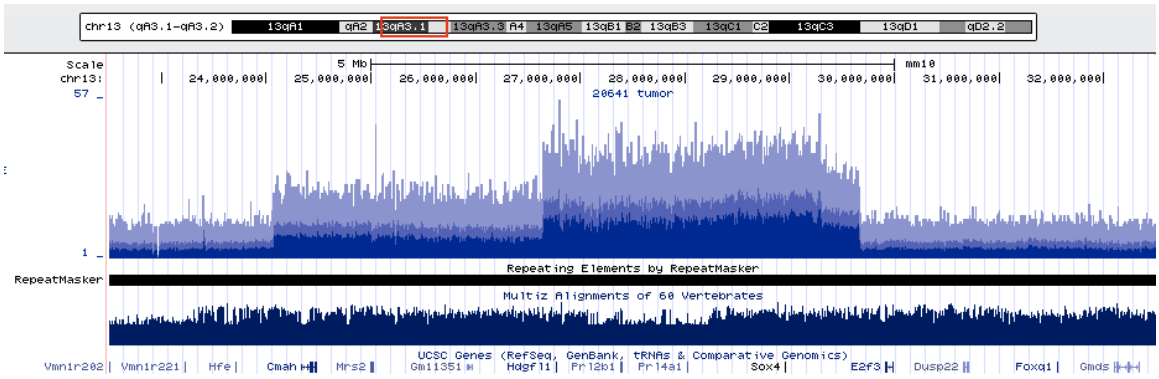

Figure 5.4. Sequencing reads that map to a 6Mb region in chromosome 13 of mouse genome mm10 are five times more abundant than reads mapping to other regions of the genome in tumor 20648. This amplified region contains the gene Sox4, a known glioblastoma oncogene. *http://genome.ucsc.edu.*

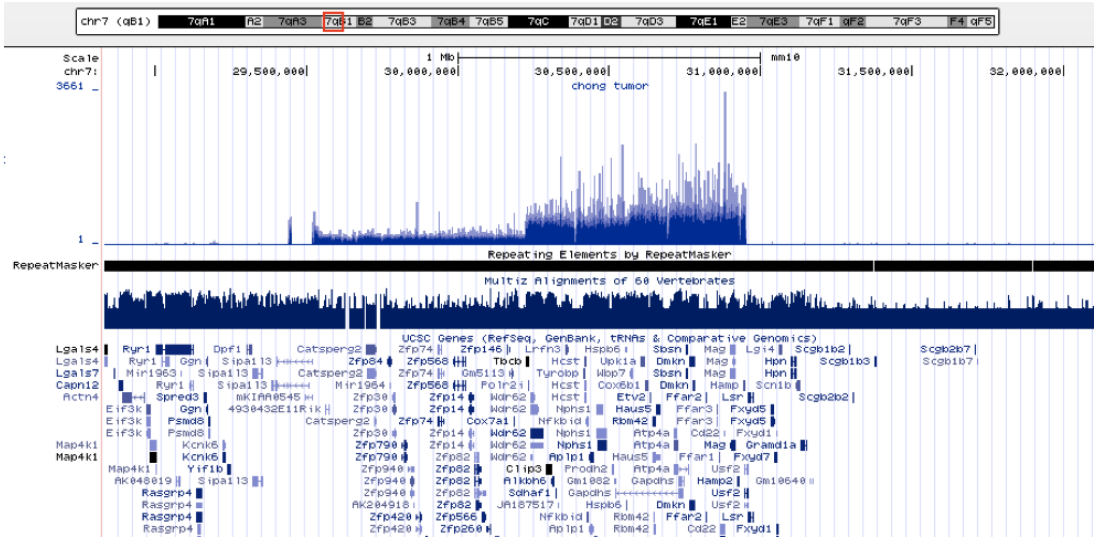

Figure 5.5. Sequencing reads that map to a 1Mb region in chromosome 7 of mouse genome mm10 are 500X more abundant than reads mapping to other regions of the genome. This amplified region contains the genes Cd22, Nfkbid, and many others. *http://genome.ucsc.edu.*

Discussion

In the p53/Nf1 MADM glioblastoma mouse model, p53/Nf1 knockout alone is insufficient to cause malignant transformation. This suggests that additional mutations are required for the mutant cells to progress into a tumor, as predicted by the mulit-hit hypothesis [16]. The intertumor heterogeneity found in different brain tumor types may be caused by different somatic mutations or developmental origins, or a combination of both [17,18]. MADM-induced knockout of p53/Nf1 results in glioma only in OPC cells, suggesting that an important factor is expressed by OPCs that is required for malignant transformation, or that an important developmental switch exists within OPC cells that can lead to glioma when improperly activated. Another possibility is that specific mutations are sufficient to cause glioma, but only if they they occur in OPCs.

Double-stranded DNA breaks and rearrangements are common in tumor genomes. Glioblastoma is known to have a higher incidence of chromothripsis, a large catastrophic genome rearrangement, compared to other tumor types [19]. However, not much is known about the extent of genome rearrangements in glioma tumors, as most studies of glioma heterogeneity focus on exome sequences or a few cancer genes [20,21]. The acute onset of typical glioma, and the higher incidence of chromothripsis suggest that genomic instability may be an important factor in glioma progression. In order to assess the genomic stability, whole genome tumor sequencing with matched nontumor samples is necessary to identify any large-scale chromosomal rearrangements not present in a wild-type genome.

The three MADM glioblastoma tumors sequenced in this study were each found to contain one or two large genomic amplifications around known cancer oncogenes. Besides the large amplifications discovered, the tumor genomes were surprisingly intact, with smooth and even genome coverage throughout. Each tumor sequenced appears to have sustained its own independent genomic lesion near an oncogene, suggesting that double-stranded breaks are a common route to tumorigenesis in this type of mouse glioblastoma. These results imply that genomic instability leads to malignant transformation in glioblastoma, and that there are several possible genetic routes leading to tumorigenesis. The lack of genomic lesions around noncancerous genes appears to suggest that once an oncogenic amplification is introduced, it is quickly selected for by the tumor.

Because the genome amplifications must have originated from double-stranded breaks, it is plausible that an upstream mutation exists which effects genome stability of OPCs. If there is indeed a SNP driving the genome amplification, it could be discovered by PELE-sequencing smaller tumors to higher depths of coverage, as described in Chapter IV. It is important to determine if such an upstream defect exists, as that would be a very attractive therapeutic target. However, it is possible that the genomic instability is simply due to random chance, in which case, the frequency of genomic amplifications in specific genes should be investigated with a larger sample size to uncover patterns of amplifications. A true understanding of the developmental and mutations origins of gliomagenesis will require a synthesis of the knowledge from developmental biology with our understanding of mutations and reorganizations that lead to malignant transformation.

CHAPTER VI

CONCLUSION

Next-Generation Sequencing (NGS) is a powerful tool that has revolutionized our understanding of genomic structure and variation for a diverse range of organisms. Unfortunately, NGS has some serious technical limitations due to the short sequencing read length and high nucleotide error rate produced by current NGS sequencing platforms such as Illumina. To address these issues, I have co-developed two new NGS techniques that improve current NGS genotyping and genome assembly, which are described in Chapters II and III. To improve genome assemblies resulting from short sequencing reads, local assembly of RAD-PE reads into contiguous sequences is employed to produce synthetic long reads for genome assembly, described in Chapter II. In order to improve the error rate of current NGS platforms, the Paired-End Low Error (PELE) Sequencing method was developed, which is a powerful method of sequencing rare genetic variants, described in Chapter III. Combined, these two new methods allow for the assessment of the total genetic potential of heterogeneous populations, including tumors, viral populations, microbiomes, and mixed pools of individuals, and for the understanding of the genomic architecture of various species.

Tumor DNA sequencing is a technically challenging pursuit due to the highly heterogeneous nature of tumors which allows them to produce a wide range of different cell types with different genomes and different levels of aggressiveness. Biopsies often give an incomplete picture of the genetic capability of a tumor, especially for advanced tumors. The driver mutations of a tumor may exist in a rare “tumor stem cell” population

that is impossible to detect with standard NGS DNA Sequencing techniques. With PELE-Seq it is now possible to know with certainty which rare variants are present in a tumor, which is crucial for an understanding of tumor initiation and progression. PELE-Sequencing can also be used to detect tumor DNA in the blood of a cancer patient, as described in Chapter IV which described the PELE-Sequencing of the blood from a human osteosarcoma patient. With PELE-Seq, rare mutations below 1% were detected in the blood DNA that were predicted to impact protein function in known cancer genes. The Mosaic Analysis with Double Markers (MADM) mouse model of glioblastoma is an extremely useful tool for investigating early tumorigenesis, because it produces glioblastoma tumors in a mosaic manner that can be identified by green fluorescent protein (GFP) expression. By sequencing the whole genomes of three p53/Nf1 KO mouse glioblastomas, two of them with matched nontumorous DNA, each tumor was found to contain one or two large genomic amplifications not present in the wildtype mice. This work was described in Chapter V. The large genomic amplifications seem to have appeared quickly and suggest that double-stranded breaks in DNA are a common route of tumor evolution in mouse p53/Nf1 KO glioblastoma.

Next-generation sequencing has opened the door to a new understanding of the variability and structure of living genomes. I have co-developed two new NGS methods that overcome the limitations of current NGS techniques: the short read length of the sequencing reads, and the high error of the nucleotide sequences generated. These techniques allow for much higher resolution when investigating genomic structure and variation, especially when sequencing heterogeneous populations of cells, such as tumors. It is now possible to know with certainty the number of polymorphisms present

in a genome, and to produce more complete genome assemblies using short sequencing reads. With these new tools to accurately assess genomes, there is much work to be done in order to uncover a functional understanding of living genomes.

REFERENCES CITED

Chapter I

1. Harismendy O, Ng PC, Strausberg RL, Wang X, Stockwell TB, Beeson KY, et al. Evaluation of next generation sequencing platforms for population targeted sequencing studies. Genome Biol. 2009;10:R32.

2. Pop M, Salzberg SL. Bioinformatics challenges of new sequencing technology. Trends Genet. 2008;24(3): 142-9.

3. Zhi D, Liu N, Zhang K. On the design and analysis of next-generation sequencing genotyping for a cohort with haplotype-informative reads. Methods. 2015; 79-80:41-46.

4. Wang Y, Lu J, Yu J, Gibbs RA, Yu F. An integrative variant analysis pipeline for accurate genotype/haplotype inference in population NGS data. Genome Res. 2013; 23(5):833-42

5. Albers CA, Lunter G, Macarthur DG, McVean G, Ouwehand WH, Durbin R. Dindel: Accurate indel calls from short-read data. 2011; Genome Res 21:961-973.

6. Li H. Improving SNP discovery by base alignment quality. Bioinformatics. 2011;27:1157–1158.

7. Orton RJ, Wright CF, Morelli MJ, King DJ, Paton DJ, King DP, et al. Distinguishing low frequency mutations from RT-PCR and sequence errors in viral deep sequencing data. BMC Genomics. 2015; 16:229.

Chapter II

1. Pop M, Salzberg SL. Bioinformatics challenges of new sequencing technology. Trends Genet. 2008;24(3): 142-9.

2. Ng PC, Kirkness EF. Whole genome sequencing. Methods Mol Biol. 2010;628: 215-26.

3. Hiatt JB, Patwardhan RP, Turner EH, Lee C, Shendure J. Parallel, tag-directed assembly of locally derived short sequence reads. Nat Methods. 2010;7(2): 119-22.

4. Li R, Fan W, Tian G, Zhu H, He L, et al. The sequence and de novo assembly of the giant panda genome. Nature. 2010;463(7279): 311-7.

5. Miller MR, Dunham JP, Amores A, Cresko WA, Johnson EA. Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers. Genome Re.s 2007;17(2): 240-8.

6. Baird NA, Etter PD, Atwood TS, Currey MC, Shiver, A. L., et al. (2008). "Rapid SNP discovery and genetic mapping using sequenced RAD markers." *PLoS ONE.* 3:e3376.

Chapter III

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

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

3. 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.

4. 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.

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

6. 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.

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

8. 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.

9. 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.

10. 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(6):1103-1112.

11. 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.

12. 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.

13. 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.

14. 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.

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

16. Schlotterer C, Kofler R, Versace E, Tobler R, Franssen SU. Combining experimental evolution with next-generation sequencing: a powerful tool to study adaptation from standing genetic variation. Heredity. 2015:114:431–440.

17. 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.

18. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, et al. The human genome browser at UCSC. Genome Res. 2002;6:996-1006.

19. Kent WJ, Zweig AS, Barber G, Hinrichs AS, Karolchi, D. BigWig and BigBed: enabling browsing of large distributed data sets. Bioinformatics. 2010;26(17): 2204-7.

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

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

Chapter IV

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

2. Hajirasouliha I, Mahmoody A, Raphael B. A combinatorial approach for analyzing intra-tumor heterogeneity from high-throughput sequencing data. Bioinformatics. 2014;30 (12):i78-i86.

3. Meacham C, Morrison S. Tumor heterogeneity and cancer cell plasticity. Nature. 2013;(501)328-337

4. Burrell R, McGranahan N, Bartek J, Swanton C. The causes and consequences of genetic heterogeneity in cancer evolution, Nature. 2013;(501)338-345

5. Gupta PB, Filmore CM, Jiang G, Shapira SD, Tao K, Kuperwasser C, Lander ES. Stochastic state transitions give rise to phenotypic equilibrium in populations of cancer cells. Cell. 2011;146(4):633-44

6. Kaur S, Singh G, Kaur K. Cancer stem cells: an insight and future perpective. J Cancer Res Ther. 2014;10(4):846-52.

7. Nguyen LV, Vanner R, Dirks P, Eaves CJ. Cancer stem cells: an evolving concept. Nat Rev Cancer. 2012; 12(2):133-43.

8. Heitzer E, Ulz P, Geigl JB. Circulating tumor DNA as a liquid biopsy for cancer. Clin Chem. 2015 Jan;61(1):112-23.

9. De Mattos-Arruda L, Weigelt B, Cortes J, Won HH, Ng CK, Nuciforo P, et al. Capturing intra-tumor genetic heterogeneity by de novo mutation profiling of circulating cell-free DNA: a proof-of-principle. Ann Oncol. 2014;25(9):1729-35.

10. Vietsch EE, van Eijck CH, Wellstein A. Circulating DNA and Micro-RNA in Patients with Pancreatic Cancer. Pancreat Disord Ther. 2015;(2)156.

11. Reva B, Antipin Y, Sander C. Predicting the Functional Impact of Protein Mutations: Application to Cancer Genomics. Nucleic Acids Research. 2011;1;39(17):e118

12. Reva BA, Antipin YA, Sander C. Determinants of protein function revealed by combinatorial entropy optimization. Genome Biol, 2007;232.

13. He Q, Kim H, Huang R, Lu W, Tang M, Shi F, et al. The Daxx/Atrx complex protects tandem repetitive elements during DNA hypomethylation by promoting H3K9 trimethylation. Cell Stem Cell. 2015;17(3):273-86

14. Langlois B, Perrot G, Schneider C, Henriet P, Emonard H, Martiny L, et al. Lrp-1 promotes cancer cell invasion by supporting ERK and inhibiting JNK signaling pathways. PLoS One. 2010;5(7):e11584

15. Poulos RC, Sloane MA, Hesson LB, Wong JW. The search for cis-regulatory driver mutations in cancer genomes. Oncotarget. 2015;PMID:26356674.

16. Melton C, Reuter JA, Spacek DV, Snyder M. Recurrent somatic mutations in regulatory regions of human cancer genomes. Nat Genet. 2015;47(7):710-6.

17. Tomasetti C, Marchionni L, Nowak MA, Parmigiani G, Vogelstein B. Only three driver mutations are required for the development of lung and colorectal cancers. Proc Natl Acad Sci USA. 2015;112(1):118-23.

18. Gatenby RA, Cunningham JJ, Brown JS. Evolutionary triage governs fitness in driver and passenger mutations and suggests targeting never mutations. Nat Commun. 2014;5:5499.

19. McFarland CD, Mirny LA, Korolev KS. Tug-of-war between driver and passenger mutations in cancer and other adaptive processes. Proc Natl Acad Sci USA. 2014 Oct 21;111(42):15138-43.

Chapter V

1. Campbell PJ, Stephens PJ, Pleasance ED, O'meara S, Li H, Santarius T. Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. Nat Genet. 2008;40(6):722-9.

2. Hillmer AM, Yao F, Inaki K, Lee WH, Ariyaratne RN, Teo AS, et al. Comprehensive long-span paired-end-tag mapping reveals characteristic patterns of structural variations in epithelial cancer genomes. Genome Res. 2011;21(5):665-75.

3. Zong H, Espinosa JS, Su HH, Muzumdar MD, Luo L. Mosaic analysis with double markers in mice. Cell. 2005;(3):479-92.

4. Chong L, Sage J, Miller MR, Roel GW, Verhaak R, Hippenmeyer S, et al. Mosaic Analysis with Double Markers (MADM) Reveals Tumor Cell-of-Origin in Glioma. Cell. 2011;146(2): 209–221.

5. Nishiyama A, Komitova M, Suzuki R, Zhu X. Polydendrocytes (NG2 cells): multifunctional cells with lineage plasticity. Nat Rev Neurosci.2009;10:9–22.

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

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

8. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. Fly (Austin). 2012;6(2): 80–92.

9. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, et al. The human genome browser at UCSC. Genome Res. 2002;6:996-1006.

10. Kent WJ, Zweig AS, Barber G, Hinrichs AS, Karolchi, D. BigWig and BigBed: enabling browsing of large distributed data sets. Bioinformatics. 2010;26(17): 2204-7.

11. Nakada M, Hayashi Y, Hamada J. Role of Eph/ephrin tyrosine kinase in malignant glioma. Neuro Oncol. 2011;13(11):1163-1170.

12. Nakada M, Anderson EM, Demuth T, Nakada S, Reavie LB, Drake KL, et al. The phosphorylation of ephrin-B2 ligand promotes glioma cell migration and invasion. Int. J. Cancer. 2010;126:1155–1165.

13. McCarthy, N. Glioblastoma: Histone mutations take the MYCN. Nature Reviews Cancer. 2013;13: 382-383.

14. Bjerke N, et al. Histone H3.3 Mutations drive pediatric glioblastoma through upregulation of MYCN. Cancer Discovery. 2013;3:512.

15. Ikushima H, Todo T, Ino Y, Takahashi M, Saito N, Miyazawa K, et al. Glioma-initiating cells retain their tumorigenicity through integration of the sox axis and Oct4 protein. Journal of Biological Chemistry. 2011;286:41434-41441.

16. Pires MM, Hopkins BD, Saal LH, Parsons RE. Alterations of EGFR, p53, and PTEN that mimc changes found in basal-like cancer promote transformation of human mammary epithelial cells. Cancer Biol Ther. 2013;14(3):246-53.

17. Liu C, Zong H. Developmental origins of brain tumors. Curr Opin Neurobiol. 2012; 22(5): 844–849.

18. Johnson RA, Wright KD, Poppleton H, Mohankumar KM, Finkelstein D, Pounds SB, et al. Cross-species genomics matches driver mutations and cell compartments to model ependymoma. Nature. 2010;466:632–636.

19. Furgason JM, Koncar RF, Michelhaugh SK, Sarkar FH, Mittal S, Sloan AE, et al. Whole genome sequence analysis links chromothripsis to EGFR, MDM2, MDM4, and CDK4 amplification in glioblastoma. Oncoscience. 2015 Jul 31;2(7):618-28.

20. Kumar, A. Boyle EA., Tokita M, Mikheev AM, Sanger MC, Girarg E. Deep sequencing of multiple regions of glial tumors reveals spatial heterogeneity for mutations in clinically relevant genes. Genome Biol. 2014 Dec 3;15(12):530.

21. Joensuu H, Puputti M, Sihto H, Tynninen O, Nupponen NN. Amplification of genes encoding KIT, PDGFRalpha, and VEGFR2 receptor tyrosine kinases is frequent in glioblastoma multiforme. J Pathol. 2005 Oct;207(2):224-31.